T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TWITTER VERİLERİ İLE TÜRK TELEVİZYONLARI İZLENME ORANI SIRALAMALARI TAHMİNİ

CENK AKARSU

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
DOÇ. DR. BANU DİRİ

İSTANBUL, 2016

ÖNSÖZ

Sosyal medya sitelerinin hızlı yükselişi ile bu siteler, büyük veri kaynakları haline gelmiştir. Kullanıcıların paylaştıkları kişisel fikirlerden oluşan bu verileri kullanarak çeşitli çıkarımlar yapılabilmektedir.

Bu çalışmada, Twitter isimli sosyal medya sitesinden alınan veriler işlenerek, Türk televizyonlarındaki programların izlenme oranı sıralamalarını tahmin edebilen bir sistem geliştirilmesi amaçlanmıştır.

Çalışma süresince bilgisi, yönlendirmeleri ve pozitif yaklaşımı ile bana büyük destek olan değerli hocam Doç. Dr. Banu Diri'ye sonsuz teşekkürlerimi sunarım.

Çalışmamda ve hayatımın her alanında desteklerini esirgemeyen aileme çok teşekkür ederim.

Şubat, 2016

Cenk AKARSU

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

F_{Ni}	Negatif tivitlerin favorilenme sayısı
F_{Pi}	Pozitif tivitlerin favorilenme sayısı
F_{PN}	Pozitif Tivit Oranı - Negatif Tivit Oranı katsayısı
F_T	Toplam Tivit Oranı katsayısı
O_i	i kullanıcısının program hakkında attığı tivit sayısının ile kullanıcının toplam tivit sayısına oranı
P	Bir yayın günü için alınan fark puanı
$P(C_i)$	C_i sınıfının olasılığı
$P(C_i \mid X)$	X örneğinin C_i sınıfından olması olasılığı
P(X)	X örneğinin olasılığı
$P(X \mid C_i)$	C_i sınıfından bir örneğin x örneği olma olasılığı
R_{Ni}	Negatif tivitlerin retweet sayısı
R_{Pi}	Pozitif tivitlerin retweet sayısı
S_i	Günlük program puanı
S_{Ni}	Programın günlük Negatif Tivit Oranı değeri
S_{Pi}	Programın günlük Pozitif Tivit Oranı değeri
S_{ri}	i. programın resmi izlenme oranı listesindeki sırası
S_{Ti}	Programın günlük Toplam Tivit Oranı değeri
S_{ti}	i. programın tahmin edilen izlenme oranı listesindeki sırası
T_i	Programın günlük toplam tivit sayısı
T_{MAX}	Hesaplama yapılan gün içinde hakkında en fazla tivit atılmış olan programın toplam tivit sayısı
T_{Ni}	Programın günlük toplam negatif tivit sayısı
T_{Pi}	Programın günlük toplam pozitif tivit sayısı
T_{Ti}	i kullanıcısının toplam tivit sayısı

KISALTMA LİSTESİ

API	Application Programming Interface
ARFF	Attribute-Relation File Format
DVM	Destek Vektör Makinesi
GPO	Günlük program puanı
JSON	JavaScript Object Notation
NB	Naive Bayes
NTO	Negatif Tivit Oranı
PTO	Pozitif Tivit Oranı
RO	Rastgele Orman
SQL	Structured Query Language
TTO	Toplam Tivit Oranı

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

TWITTER VERİLERİ İLE TÜRK TELEVİZYONLARI İZLENME ORANI SIRALAMALARI TAHMİNİ

Cenk AKARSU

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Banu DİRİ

Sosyal Web Madenciliği, günümüzde oldukça ilgi çekmekte olan bir alandır. İnsanlar tarafından dinamik olarak oluşturulan ve büyük bir hızla büyüyen sosyal medya verileri, grip salgınları ve seçim sonuçları gibi birçok farklı konuda önceden tahminlerde bulunmak için kullanılmıştır. Bu çalışmada da oldukça zengin bir veri kaynağı olan Twitter isimli sosyal medya platformundan toplanan veriler ile Türk televizyonlarında yayınlanan programların izlenme oranı sıralaması tahminlerini yapmak amacıyla bir sistem geliştirilmiştir. Bu sistem; otomatik olarak çeşitli kaynaklardan veri toplama, verileri ilişkisel olarak depolama, veriler üzerinde temizlik, doğal dil işleme, anlamsal sınıflandırma ve izlenme oranı sıralaması tahminleri yapabilme yeteneklerine sahiptir.

Sistem geliştirilirken Naive Bayes, Destek Vektör Makinesi ve Rastgele Orman sınıflandırma algoritmalar, Twitter'dan alınan ve sistem tarafından çeşitli şekillerde işlenen veriler ile oluşturulan eğitim setleri kullanılarak eğitilmiş ve test edilmiştir. Çalışmada testlerin karşılaştırmalı sonuçları verilmiştir. Testlerde en yüksek başarıyı elde eden sınıflandırma algoritması sistemde kullanılarak, televizyon programları hakkında toplanan veriler sınıflandırılmıştır. Bu sınıflandırma sonuçları kullanılarak her program için gün bazında izlenme oranı puanları hesaplanmış ve bu puanlara göre izlenme oranı sıralaması tahminleri yapılmıştır.

Sistem tarafından tahmin edilen ve resmi sonuçlar, farklı kriterlere göre ve farklı yöntemler kullanılarak karşılaştırılmıştır. Bu karşılaştırma sonuçları paylaşılmıştır.

Anahtar Kelimeler: İzlenme oranı, doğal dil işleme, sınıflandırma, Twitter, sosyal medya

ABSTRACT

PREDICTING TURKISH TELEVISION RATINGS RESULTS USING TWITTER DATA

Cenk AKARSU

Department of Computer Engineering

MSc. Thesis

Adviser: Assoc. Prof. Dr. Banu DİRİ

At the present time, Social Web Mining is a field that attracts a lot of attention. The dynamic human data, which grows quite fast, has been used to predict a variety of things like flu trends and political election results. In this Project, a system has been developed to predict the rating results of Turkish TV programs using Twitter, a rich source of social data. This system has abilities like data collection, storing data relationally, data cleaning, natural language processing, machine learning and rating result prediction.

In the process of system development, Naive Bayes, Support Vector Machine and Random Forest classification algorithms were trained and tested using the training data that created with the tweet collected from Twitter and processed by the system in various ways. The comparative test results were given. The data that has been collected for television programs were classified using the most successful classification algorithm, based on the tests. Using these classification results, daily rating scores were calculated for television programs and rating lists were predicted based on these scores.

The results predicted by the system and the official results were compared by various criterias and using various methods. The comparison results were given.

Keywords: Rating, natural language processing, classification, Twitter, social media

BÖLÜM 1

GİRİŞ

Günümüzde sosyal medya platformlarının hızlı yükselişi ile bu platformlardaki kullanıcı sayısı ve bu kullanıcılar tarafından yapılan paylaşımların miktarında büyük bir artış meydana gelmiştir. İnsanların, hemen hemen her konuda kişisel fikir ve düşüncelerini paylaştıkları bu platformlar, önemli birer veri kaynağı haline gelmiştir. Öyle ki, insanların ruh hallerini ölçmek için bu veriler kullanılabilmektedir [1]. Hatta sosyal medya üzerinden yapılan imaj çalışmaları ile insanların algılarının olumlu ya da olumsuz şekilde yönlendirmek bile mümkündür. Bu yüzden, günümüzde birçok firma, kurum, kişi tarafından sosyal medya üzerinde imaj çalışmaları ve algı ölçümleri yapılmaktadır. Örneğin, yeni bir ürün piyasaya süren bir firma, müşterilerin bu ürünü beğenip beğenmediğini sosyal medyadaki; herhangi bir aracı olmadan, direkt olarak müşteriler tarafından oluşturulmuş olan verileri inceleyerek öğrenebilmektedir. Sosyal medya üzerinden yapacağı bir kampanya ile insanların ürüne olan talebini de arttırabilmektedir. Bu nedenlerden dolayı günümüzde birçok firma tarafından sosyal medya üzerinden yapılan reklam kampanyalarına televizyon ya da gazete reklamlarından daha fazla önem verilmektedir. Eskiden çoğunlukla anketler ile elde edilen kullanıcı görüşlerinin, kullanıcılar tarafından gönüllü olarak sosyal medyada paylaşılması ile firmalar tarafından daha çabuk ulaşılabilir hale gelmesi, firmaların kullanıcı isteklerine göre ürün ve hizmet geliştirmesini kolaylaştırmıştır. Bunların dışında sosyal medya verilerini kullanarak çeşitli konularda akademik çalışmalar da yapılmaktadır. Veri miktarının büyüklüğü ve çeşitliliği birçok araştırmacının ilgisini çekmektedir.

Sosyal medyada kullanıcı görüşlerinin etkin olarak takip edildiği sektörlerden biri de televizyon sektörüdür. Sosyal medyanın düşünceleri hızlıca paylaşmaya olanak vermesi nedeniyle insanlar televizyonda izledikleri programlar hakkındaki görüşlerini anlık olarak paylaşabilmektedirler. Öyle ki, yayınladıkları programlar hakkındaki paylaşımları takip eden kanallar, program içeriklerini kullanıcı isteklerine göre şekillendirebilmektedir. Örneğin sosyal medyada çok sevilen karakterlerin, dizilerdeki süreleri artırılmaktadır. Benzer şekilde hakkında çok fazla olumsuz paylaşım bulunan programlar da yayından kaldırılabilmektedir. Kanallar, kullanıcılar tarafından yapılan paylaşımları birer talep olarak ele alıp çoğunluğun istediği talepleri gerçekleştirerek izlenme oranlarını arttırmaya çalışmaktadır.

Sosyal medya verileri ile insanların çeşitli konulardaki düşünceleri ölçülebilir fakat bu verilerin miktarı çoğu zaman gözle kontrol edilip anlamlandırılamayacak kadar büyüktür. Bu yüzden sosyal medya verileri üzerinde algı analizi yapılması ile ilgili çalışmalar yürütülmektedir. Bazı çalışmalarda sosyal medya verileri kullanılarak sadece mevcut düşünceler değil, geleceğe yönelik çeşitli çıkarımlar da elde edilebilmektedir. Twitter üzerinden alınan veriler kullanılarak, toplumun duygusal durumu ile borsadaki değişikliklerin ilişkili olduğunu ve sosyal medya verileri kullanılarak önceden çıkarımlar yapılabileceğini gösteren çalışma [2] bunlara örnek olarak verilebilir.

Sosyal medya verilerini otomatik olarak anlamlandırabilmek ve çıkarımlar yapabilmek için bu verilerin çeşitli işlemlerden geçirilmesi gerekmektedir. Kullanılacak olan verilerin toplanması, temizlenmesi, düzeltilmesi ve sınıflandırılması işlemlerini yapabilecek olan sistemlerin geliştirilmesi gerekmektedir. Bu çalışmada da sosyal medya sitesi Twitter'dan elde edilen Türkçe tivitler kullanılarak, televizyon programlarının izlenme oranı sıralamalarını önceden tahmin edebilecek bir sistem geliştirilmesi hedeflenmiştir.

1.1 Literatür Özeti

Sosyal medya verileri üzerinde sınıflandırma ve verileri kullanarak çıkarımlar yapmak amacıyla yapılmış çok sayıda çalışma mevcuttur. Sosyal medya platformlarından Twitter, hızla büyüyen ve kolay ulaşılabilen verileri sayesinde birçok çalışma tarafından tercih edilmiştir.

“Twitter mood predicts the stock market” [2] isimli çalışmada 28-02-2008 / 19-12-2008 tarihleri arasında, yaklaşık 2.700.000 kullanıcı tarafından paylaşılmış olan 9.853.498 tivit kullanılarak Dow Jones Borsası Endüstri Endeksi’nde meydana gelen değişimlerin tahmini yapılmıştır. Çalışmada, yazarın ruh hali ile ilgili bilgi veren tivitler kullanılarak toplumun genel ruh halini ölçmek amaçlanmış ve bu ruh hali ile ilgili borsadaki iniş ve çıkışların ilişkisi araştırılmıştır. Sonuç olarak %87,6’lık bir başarı ile Dow Jones Borsası Endüstri Endeksi’nde gerçekleşen günlük iniş ve çıkışlar tahmin edilebilmiştir. Benzer şekilde, “Ant colony based approach to predict stock market movement from mood collected on Twitter” [3] isimli çalışmada da Twitter’dan elde edilen ruh hali bilgisi ile borsa tahmini yapılmaya çalışılmıştır. “Twitter Sentiment Classification Using Machine Learning Techniques for Stock Markets” [4] isimli çalışmada ise Twitter’dan alınan veriler üzerinde algı analizi yapılarak veriler olumlu, olumsuz ya da nötr olarak sınıflandırılmış ve elde edilen sonuçlara göre 4 teknoloji şirketinin (Twitter, Google, Facebook ve Tesla) borsa tahminleri yapılmıştır.

“Predicting the Future With Social Media” [5] isimli çalışmada, Huberman ve Asur tarafından Twitter verileri, sinema filmlerinin gişede elde edeceği gelirleri filmler henüz vizyona girmeden tahmin etmek için kullanılmıştır. Twitter’dan veri alınırken anahtar kelime olarak film isimleri kullanılmış, 3 aylık bir zaman aralığı boyunca 24 farklı filmle ilgili 2.890.000 tivit toplanmıştır. Tivitler kullanılarak yapılan tahminler, Hollywood Stock Exchange adlı siteden alınan veriler ile karşılaştırılmıştır. Benzer bir çalışma da “Prediction of Movies Box Office Performance Using Social Media” [6] isimli çalışmadır. Bu çalışmada Twitter’ın yanı sıra IMDB ve Youtube’dan alınan veriler de kullanılmıştır.

“Towards detecting influenza epidemics by analyzing Twitter messages” [7] isimli çalışmada, Twitter üzerinden alınan, 10 haftalık bir zaman aralığını kapsayan 500.000’den fazla tivit kullanılarak grip salgınları tespit edilmeye çalışılmıştır. Basit doğrusal regresyon ve çoklu doğrusal regresyon kullanılan bu çalışmada elde edilen veriler, ABD Hastalık Kontrol ve Korunma Merkezleri’nden alınan veriler ile %78 oranında benzerlik göstermiştir. Benzer bir çalışma da “Predicting Flu Trends using Twitter Data” [8] isimli çalışmadır. Bu çalışmada da grip ile ilgili paylaşılmış olan tivitler kullanılarak grip salgınları bulunmaya çalışılmıştır.

“Twitter as a Tool for Predicting Elections Results” [9] isimli çalışmada, geliştirilen Taratweet isimli araç kullanılarak 2011 ve 2012 yıllarında İspanya’da yapılan 3 farklı seçim sonuçları tahmin edilmeye çalışılmıştır. Yerel seçimler için yapılan deney 105.282; genel seçimler için yapılan deney 259,016; Endülüs’te yapılan seçim için yapılan deney 176.000 adet tivit kullanılarak gerçekleştirilmiştir. Çalışmada siyasi partilere verilen oylar ile bu partilerin Twitter’da bahsedilme oranları karşılaştırılmış ve uyumlu oldukları görülmüştür. Benzer şekilde [10] de Endonezya başkanlık seçimi; [11] de Hindistan genel seçimi ve [12] de Pakistan’da yapılan seçim sonuçları Twitter kullanılarak tahmin edilmeye çalışılmıştır.

“Twitter ile TV Program Reytinglerinin Belirlenmesi” [13] isimli çalışmada, Twitter’dan elde edilen veriler, içerdikleri kelimelere göre olumlu, olumsuz ya da nötr şeklinde 3 farklı sınıfa ayrılmış ve hesaplanan sonuç değerine bağlı olarak televizyon izlenme oranı sonuçlarının tutarlılığı test edilmeye çalışılmıştır.

1.2 Tezin Amacı

Kolay kullanımı ve paylaşımları hızlı bir şekilde birçok kişiye ulaştıran tasarımı sayesinde Twitter, insanların kendini ifade etmek için kullandığı ortamlardan biri haline gelmiştir. Aynı zamanda birçok firma ve ünlü kişinin de birer kullanıcı olarak yer aldığı bu platform, kullanıcıların fikirlerini, şikâyetlerini, övgülerini ve sorularını istedikleri kişiye çabucak ulaştırıp, cevap alabildiği bir iletişim aracı olarak da görev yapmaktadır.

Her konuda olduğu gibi, televizyon programları hakkındaki düşünceler de Twitter üzerinde kullanıcılar tarafından paylaşılmaktadır. Bu veriler incelenerek, yayınlanan programlar hakkında izleyicilerin algılarını öğrenmek mümkündür. Fakat tüm sosyal medya platformlarında olduğu gibi Twitter’da da, kullanıcılar tarafından oluşturulan veriler oldukça kirli, tekrarlayan ve sınıflandırılmamış biçimdedir. Bu veriler ile izleyici algısının öğrenilebilmesi için öncelikle verilerin işlenmesi gerekmektedir.

Bu çalışmada Twitter verileri doğal dil işleme, veri madenciliği ve makine öğrenmesi teknikleri kullanılarak işlenecek ve elde edilecek izleyici görüşlerine göre televizyon programları izlenme oranı sıralaması tahminleri yapılması amaçlanmıştır. Bunun için öncelikle Türk televizyonlarında yayınlanan programlar ve bu programlar hakkında

paylaşılan tivitler toplanmış ve veri tabanına kaydedilmiştir. Bu veri toplama işlemi, Bölüm 3.1'de detaylandırılmıştır.

Twitter üzerinden alınan veriler çok miktarda tekrarlı biçimdedir. Ayrıca anahtar kelimeler ile erişilebilen bu veriler, her zaman barındırdığı anahtar kelime ile ilgili yazılmamış olabilmektedir. Çalışmanın ikinci adımda, sistem tarafından hangi tivitlerin hesaba katılacağını belirlemek için bir eleme işlemi yapılmıştır. Bölüm 3.3'te bu işlemlerin detayları paylaşılmıştır.

Toplanan veriler oldukça karmaşık durumdadır ve ham haliyle kullanılmaları zordur. Bu yüzden veriler içindeki yazım hataları, Twitter ve sosyal medya jargonuna ait sözcükler işlenmeli ve tivitler sistem tarafından anlaşılabilir bir hale getirilmelidir. Bu yüzden çalışmanın üçüncü adımında veriler üzerinde doğal dil işleme yapılarak bu temizlik ve düzeltme işlemleri gerçekleştirilmiştir. Bu işlemlerin detayları Bölüm 3.4'te anlatılmaktadır.

Çalışmanın dördüncü adımında, işlenmiş olan tivitler makine öğrenmesi algoritmaları kullanılarak anlamsal olarak sınıflandırılmış ve etiketlenmiştir. Beşinci adımda ise sınıflandırılmış olan bu verilerin etiketlerine göre hesaplamalar yapılarak, gün bazında izlenme oranı sıralaması tahminleri yapılmıştır. Bölüm 3.5 ve Bölüm 3.6, bu işlemlerden bahsetmektedir.

Çalışmanın son adımında ise sistem tarafından tahmin edilen sıralamalar ile resmi izlenme oranı sonuçları çeşitli kriterlere göre karşılaştırılmış ve başarı ölçümü yapılmıştır. Sonuçlar Bölüm 5'te paylaşılmıştır.

1.3 Hipotez

Bu çalışmada, Twitter'dan alınan veriler ile resmi izlenme oranı sonuçlarının önceden tahmin edilebilmesi amaçlanmıştır. Bunun için öncelikle Twitter'dan televizyon programları ile ilgili olan tivitler çekilmiştir ve bu tivitler elemeden geçirilip, doğal dil işleme yöntemleri ile işlenip, makine öğrenmesi yöntemleri ile sınıflandırıldıktan sonra izlenme oranı sıralaması listeleri tahmininde kullanılmıştır.

İkinci bölümde Twitter ve Twitter verileri ile ilgili bilgiler verilmiştir. Üçüncü bölümde geliştirilen sistem ve çalışma yöntemlerinden bahsedilmiştir. Dördüncü bölümde

oluşturulan eğitim setleri hakkında bilgiler verilmiş; beşinci bölümde ise deneysel sonuçlar paylaşılmıştır.

BÖLÜM 2

TWITTER VERİLERİ

Bu bölümde Twitter, Twitter verileri ve sistem için oluşturulan eğitim seti hakkında bilgiler verilmiştir. Bölüm 2.1'de Twitter hakkında genel ve istatistiki bilgiler verilmiş ve sitenin işleyişi anlatılmıştır. Bölüm 2.2 ise tivitler ve tivit yazılırken kullanılan bazı kısaltma yöntemleri hakkında bilgiler vermektedir.

2.1 Twitter Nedir?

Twitter, 2006 yılında Jack Dorsey tarafından kurulmuş olan bir mikroblog sitesi ve sosyal ağdır. Kullanıcılarının en fazla 140 karakterden oluşan ve tivit adı verilen mesajlar paylaşmalarına olanak sağlayan bu site günümüzde en çok ziyaret edilen web sitelerinden biridir. Tivitler, resim ya da videolar da barındırabilmektedir. 30 Eylül 2015 sonu itibarıyla Twitter hakkındaki bazı kullanım bilgileri şu şekildedir [14]:

- Aylık 320 milyon aktif kullanıcı sayısına sahiptir.
- Entegre edilmiş tivit içeren siteler aylık 1 milyar eşsiz ziyaret almaktadır.
- Mobil cihazlarda aktif kullanıcı oranı %80'dir.
- 35'den fazla dil desteklemektedir.
- ABD dışında bulunan hesapların oranı %79'dur.

Twitter'da, kullanıcıların birbirlerinin hesaplarını takip etmesine dayanan bir sistem mevcuttur. Site tarafından kullanıcılara, takip ettikleri kişilerin paylaştıkları tivitler ana sayfada gösterilmektedir. Kullanıcılar, istedikleri takdirde hesaplarını koruma altına alabilirler. Korunan hesaplardan paylaşılan tivitleri sadece o hesabın takipçileri

tarafından görülebilir. Hesabın takipçisi olmak için de hesap sahibinin onayı gereklidir. Koruma altına alınmamış olan hesaplardan paylaşılan tivitler ise o hesabın profil sayfasına giren herkes tarafından görüntülenebilmektedir. Tivitleri görüntülemek için kullanılan bir yöntem de Twitter tarafından sağlanan arama özelliğini kullanmaktır. Bu özellik, kelime, kullanıcı adı, lokasyon, tarih gibi kriterler ile kullanılabilmektedir. Arama sonuçları en yeni tarihli tivitten başlayarak, eskiye doğru olacak şekilde gösterilir.

Twitter'da kullanıcılar, tivit paylaşmak dışında, başkaları tarafından paylaşılmış olan tivitleri retweet edebilirler veya favorileyebilirler. Retweet etme işlemi, tivitin altındaki retweet butonuna basılarak, başkasına ait bir tivitin retweet eden kullanıcının profilinde de görünmesini sağlayan işlemdir. Favorileme ise tivitin altındaki favori butonuna basılarak, tivitin favorileyen kullanıcının favoriler listesinde görünmesini sağlayan işleme denir. Kullanıcılar beğendikleri tivitleri retweet ederek ya da favorileyerek takipçilerine paylaşmış olurlar.

Twitter, anlık olarak en çok 10 adet konuyu listelediği ve "Gündem" adı verdiği listeler barındırır. Bu listeler tivitler içinde en çok kullanılan sözlerden oluşur. Şehir bazında, ülke bazında ve dünya genelinde en çok konuşulan konuları gösteren listeler mevcuttur. Kullanıcılar aynı sözleri barındıran çok sayıda tivit atarak bu listeye girip, dikkat çekmeye çalışır.

2.2 Tivit İçerikleri

Tivit başına 140 karakter sınırı bulunmaktadır. Bu nedenle kullanıcılar daha az karakter ile daha çok fikir paylaşabilmek için çeşitli kısaltma yöntemleri kullanmaktadır. "Etiket" olarak adlandırılan ve diyez (#) karakteri ile başlayan kelime veya söz öbeklerinin tivit içeriğine yazılması bu yöntemlerden biridir. Örneğin:

"*Gönüllüler oyunları kaybedeli beri programdan soğudum #SurvivorAllStar*"

Etiketler, aynı konulardan bahseden tivitleri bir araya toplamak için de kullanılır. Kullanıcılar, bir tivit içindeki etikete tıkladıklarında, o etiketi barındıran tivitler toplu olarak gösterilir. Günümüzde birçok televizyon programı, yayınlanan bölüm için oluşturdukları etiketleri paylaşıp, izleyicileri bu etiketleri barındıran tivitler yazmaya yönlendirmektedir. Amaç etiketi Gündem'e sokarak reklam yapmaktır.

Etiketler dışında, kelimeleri eksik harfler ile yazmak da oldukça sık başvurulan bir yöntemdir. Örneğin, “kanal” yerine “knl”, “program” yerine “prgrm” yazılarak, daha az harf ile daha fazla şey anlatılabilmektedir. Ayrıca, Türkçe kelimeler yerine yabancı dilde kullanılan kısaltmalar kullanmak da kullanıcılar tarafından sıkça kullanılır. Örneğin, “gülmek” yerine “lol”, “direkt mesaj” yerine “dm” yazmak gibi.

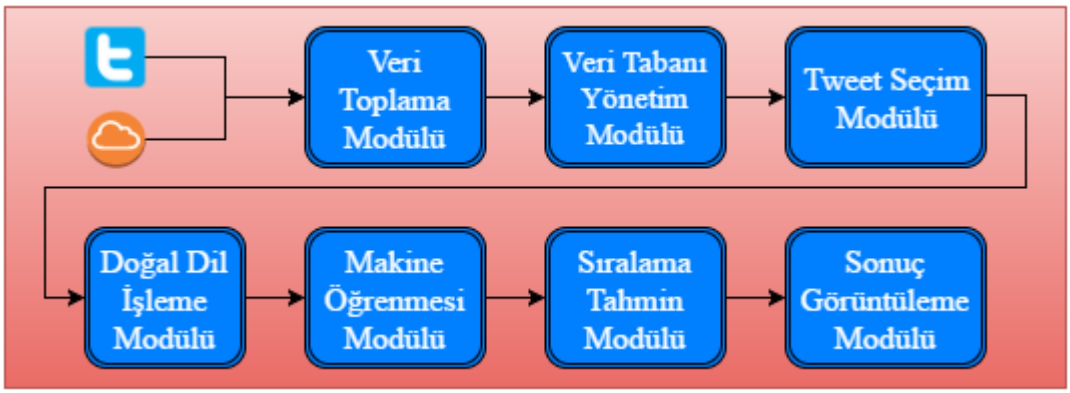
BÖLÜM 3

GELİŞTİRİLEN SİSTEM

Bu çalışmada geliştirilen sistem, birden fazla modülün birleşiminden oluşmaktadır. Şekil 3.1'de görülebilen bu modüller,

- Veri Toplama Modülü,
- Veri Tabanı Yönetim Modülü
- Tivit Seçim Modülü
- Doğal Dil İşleme Modülü
- Makine Öğrenmesi Modülü
- Sıralama Tahmin Modülü
- Sonuç Görüntüleme Modülü olmak üzere 7 adettir.

Bu bölümde sistemin modülleri tanıtılıp, çalışma mantıkları anlatılacak ve sistemdeki veriler ile ilgili bilgi verilecektir.



Şekil 3. 1 Geliştirilen sistem modülleri

3.1 Veri Toplama Modülü

Veri Toplama Modülü, web ortamından sistem için gereken verilerin toplanıp, uygun formata çevrilerek kaydedilmesinden sorumlu olan modüldür. Visual Studio ortamında C# dili kullanılarak geliştirilmiştir. Bu modül, 2 farklı kaynaktan, 2 farklı türde veri toplamaktadır. Bu veri türlerinden ilkini Twitter'dan alınan tivitler oluşturmaktadır. Diğer veri türü ise, günlük olarak açıklanan ve internette yayınlanan Türk televizyon kanallarına ait izlenme oranlarından elde edilmektedir. Modül belirli zaman aralıkları ile çalışarak gerekli kaynaklardan veri toplamaktadır. Bu bölümde tivitlerin ve izlenme oranı bilgilerinin toplanması, sistem tarafından anlaşılır şekle dönüştürülmesi ve saklanması anlatılmaktadır.

3.1.1 Tivitlerin Alınması

Veri Toplama Modülü, kullanıcı tarafından belirlenen zaman aralıkları ile periyodik olarak tivitleri toplar ve sisteme kaydeder. Modül, tivitlere ulaşmak için Twitter API kullanmaktadır. Sistemde televizyon programları bazında kaydedilmiş olan arama terimlerini kriter olarak kullanarak Twitter API içindeki tivit arama (search/tweets) servisine sorgular gönderir. Yapılan sorgulara cevap olarak gelen veriler, sistemin anlayacağı biçime getirilip, arama terimi, program, program türü, kanal vb. ilişkilere göre gereken şekilde kaydedilmektedir.

3.1.1.1 Twitter API ile Tivit Arama

Twitter API [15], kullanıcıların site dışındaki yazılımlardan da tivitlere ulaşmalarını sağlamak amacıyla Twitter tarafından yayınlanmış bir servistir. Tivit arama servisi ise Twitter API içinde bulunan, farklı parametreler ile çağrılabilen ve yapılan çağrılara JSON (JavaScript Object Notation) formatında yanıtlar dönen bir servistir. Dönen yanıtlar tivit detayları ve tiviti paylaşan kullanıcı bilgileri olmak üzere 2 ana bölümden oluşmaktadır. Tivitler, en yeni tarihliden başlayarak geriye doğru sıralanmış bir şekilde gelmektedir. Servisten dönen JSON veriler içinde birçok alan bulunmaktadır. Şekil 3.2, örnek bir tivite ait, JSON formatındaki veriyi göstermektedir.

```
{
    "coordinates": null,
    "favorited": false,
    "truncated": false,
    "created_at": "Mon Sep 24 03:35:21 +0000 2012",
    "id_str": "250075927172759552",
    "entities": {
        "urls": [],
        "hashtags": [ {"text": "freebandnames","indices": [20, 34]} ], "user_mentions": [] },
    "in_reply_to_user_id_str": null,
    "contributors": null,
    "text": "Aggressive Ponytail #freebandnames",
    "metadata": {
        "iso_language_code": "en",
        "result_type": "recent"
    },
    "retweet_count": 0,
    "in_reply_to_status_id_str": null,
    "id": 250075927172759552,
    "geo": null,
    "retweeted": false,
    "in_reply_to_user_id": null,
    "place": null,
    "in_reply_to_screen_name": null,
    "source": "",
    "in_reply_to_status_id": null
}
```


Şekil 3. 2 Örnek json veri formatı

Çalışmada bu alanların hepsine ihtiyaç duyulmadığından sadece belli alanlar seçilip sisteme kaydedilmektedir. Bu işlem Veri Toplama Modülü tarafından yapılmaktadır. Sisteme kaydedilen alanlar şunlardır:

- **id:** Tivit ID
- **text:** Tivit metni
- **favorite_count:** Tivitin favorilenme sayısı
- **retweet_count:** Tivitin retweet edilme sayısı
- **lang:** Tivit dili
- **created_at:** Tivitin oluşturulma tarihi

Arama servisinden her bir tivitle beraber, tiviti paylaşan kullanıcıya ait bilgiler de gelmektedir. Kullanıcı bilgilerinden seçilerek sisteme kaydedilen alanlar şu şekildedir:

- **id:** Kullanıcı ID
- **name:** Kullanıcının tam adı
- **screen_name:** Kullanıcı adı

- **description:** Kullanıcı hesabı için girilen tanımlama metni
- **followers_count:** Kullanıcı hesabının takipçi sayısı
- **location:** Kullanıcı hesabı için girilmiş olan lokasyon değeri
- **statuses_count:** Kullanıcının paylaşılan toplam tivit sayısı
- **friends_count:** Kullanıcı tarafından takip edilen hesapların sayısı
- **created_at:** Kullanıcı hesabının oluşturulma zamanı

Twitter API kullanımı belli limitlere sahiptir. Limitler, API içindeki servislere göre değişse de 15 dakikalık aralıklar bazındadır. Kullandığımız arama servisi için her bir geliştirici hesabı başına 15 dakikada en fazla 180 sorgu limiti vardır. Sistem için saatte 180 sorgu ile elde edilebilecek veriler çok yetersiz olacağından sistemde birden fazla geliştirici hesabı kullanarak sorgular yapma yoluna gidilmiştir. Veri Toplama Modülü yapılan her sorgu ve kullanıcı bilgisini hesaplayarak, saatlik limiti dolmamış olan kullanıcılar ile sorgulama yapar. Limiti dolan hesap beklemeye alınarak boşta olan hesaplardan biri kullanılır. Böylece limitlere takılmadan veri alınabilmektedir.

3.1.1.2 Anahtar Kelimeler

Veri Toplama Modülü, Twitter API içindeki tivit arama servisine sorgular yaparken sistemde kayıtlı olan arama terimlerini kullanmaktadır. Bu kelimeler program bazında tutulmaktadır ve şu tiplerde olabilir:

- **Program İsmi:** Televizyon programının isminden oluşan arama terimleri.

  Ör: Beyaz Show, Kim Milyoner Olmak İster
- **Oyuncu İsmi:** İlgili programdaki bir oyuncunun ya da karakterin ismi.
- **Etiket:** Programın isminden oluşturulmuş olan etiketler. Ör: #beyazshow
- **Çoklu Kelimeler:** Programın isminin, çoğunlukla program kast edilmeden kullanılan kelime ya da kelimelerden oluştuğu durumda, ismin yanına başka kelimeler (program türü, kanal ismi vb.) eklenerek oluşturulan arama terimleridir. Ör: paramparça dizi, paramparça star

Anahtar kelimeler birkaç şekilde oluşturulabilir. Sisteme her yeni bir program kaydedildiğinde, program isminden otomatik olarak Program İsmi ve Etiket türünde arama terimleri oluşturulur. Kullanıcı tarafından elle giriş de yapılabilir. Bunların dışında sistemin elde ettiği tivitleri işleyerek oluşturduğu arama terimleri de vardır. Bu arama terimleri Etiket türünde olup, genellikle programın bir bölümü süresince kullanılırlar. Günümüzde başta diziler olmak üzere birçok televizyon programı, her bölüm için yeni bir etiket belirleyerek izleyicileri bu etiketi barındıran tivitler paylaşmaya zorlamaktadır. Bu etiketleri sistemin otomatik olarak yakalaması için şu işlemler yapılmaktadır:

1. İlgili yayın günü için program hakkında en az 100 adet tivit toplanır.

2. Elde edilen tivitler içindeki etiketler bulunur ve her etiket için etiketin bulunduğu tivitler sayılır.

3. Eğer etiketin bulunduğu tivit sayısı, program için o gün toplanan tivitlerin sayısının %30'undan büyük ya da eşitse bu etiketin o gün yayınlanan diğer programlar için toplanan tivitlerin kaçında geçtiği sayılır.

4. Eğer etiket en yüksek oranla, ilgili programa ait tivitler içinde bulunuyorsa otomatik oluşturulmuş bir arama terimi olarak kaydedilir ve tivit toplamada kullanılmaya başlanır.

3.1.2 İzlenme Oranlarının Alınması

Veri Toplama Modülü günde bir kez http://www.sacitaslan.com/ adresindeki web sitesine gider ve bir önceki gün için açıklanan izlenme oranı sonuçlarını alır. Bu web sitesinde bağlantı adresinin sonuna istenen tarih yazılarak gidildiğinde, o güne ait sonuçlar listelenmektedir. Sistemin kullandığı adres formatına örnek olarak:

"http://www.sacitaslan.com/rating/gunluk/02.02.2016"

Modül tarafından oluşturulan adrese gidilerek sayfanın HTML kodu elde edilir. Bu kod içindeki alanlar modül tarafından bilinmektedir. HTML kodları işlenerek, gereken alanlardaki bilgiler ayrıştırılır ve sisteme kanal, program ve izlenme oranı tipi bazında kaydedilir.

3.2 Veri Tabanı Yönetim Modülü

Veri Tabanı Yönetim Modülü, sistemde tutulan bilgilerin saklandığı ve sorularla alınıp işlendiği bölümdür. Bu bölüm bir Microsoft SQL SERVER 2014 veri tabanından oluşmaktadır. Verilerin farklı tablolarda ve birbirleri ile ilişkili olarak tutulması gerektiğinden bu sistem tercih edilmiştir. Tutulan veriler Transact-SQL dili ile ilişkisel olarak sorgulanıp kullanılmıştır. Modül içindeki tablolar şu bilgileri tutmaktadır:

- Tivitler
- Twitter kullanıcıları
- Kanallar
- Kanal türleri
- Programlar
- Program türleri
- Anahtar kelimeler
- Anahtar kelime türleri
- Açıklanan izlenme oranları
- Hesaplanmış program puanları
- Günlük sıralama tahminleri

3.3 Tivit Seçim Modülü

Sistem, izleyici düşüncelerini yansıtan tivitleri işleme alarak tahminde bulunmaya çalışır. Fakat toplanan tivitlerin hepsi, aslında ilişkili olarak çekildikleri programdan bahsetmiyor olabilir. Özel hazırlanmış programlar tarafından otomatik olarak paylaşılmış tivitler de mevcuttur. Bu tivitler bazı günlerde toplanan tivitlerin çoğunluğunu oluşturmaktadır. Bu yüzden kullanılacak tivitleri belirlemek için bir seçim işlemi yapılmalıdır. Tivit Seçim Modülü, sınıflandırma işlemi öncesinde, izlenme oranı tahmini üzerinde bir etkisi olmayan bu tivitlerin tanınıp, ayrıştırıldığı bölümdür. Ayrıştırılan tivitler genel olarak şu türlerde olmuştur:

- Dizi, film vb. yayınlayan siteler ya da yayıncı kanallar tarafından otomatik olarak oluşturulup atılan tivitler

  Ör: "*#atvAnaHaber başlıyor! http://t.co/lOgJc5IucN"*

- Tivit içinde bulunan etiketlerden gündem oluşturmak için otomatik olarak çok miktarda atılan, aynı metne sahip tivitler

  Ör: *"Atama için 'Haydi İnşallah!': https://t.co/MPxpmuNHq3 #BazenDiyorumKi #AlexeVeda #SurvivorAllStar #40BinOgretmenNisandaistihdamaKararlı"*

- Başka konulardan bahsettiği halde görünürlüğü arttırmak amacıyla, işlem yapılan etiketlerden eklenmiş olan tivitler

  Ör: *"@poyrazkarayeldz Poyraz Karayel Twitter Türk Bot Takipci istersen Sitemize Bekleriz http://t.co/MIWYdoq7lR http://t.co/T0BFbsOGlO 836"*

Seçim işlemleri tivit kaynağına, formatına, içeriğine (tarih, link, 'bölüm, izle' vb. kelimeler) ve kullanıcı bilgilerine (toplanan tivit sayısının toplam tivit sayısına oranı, kullanıcı oluşturulma tarihi, toplanan tivitlerdeki etiket içerikleri) bakılarak yapılmaktadır. Bölüm 3.3.1'de bu işlemler anlatılmaktadır.

3.3.1 Tivitlerin Seçimi

Seçim işlemi yayın günü ve program bazında yapılmaktadır. İşlem, birkaç aşamada gerçekleştirilir.

Tivit seçim işleminin ilk adımı olarak toplanan tivitlerin kaynak bilgisine bakılır. Twitter, farklı kaynaklardan tivit paylaşmaya olanak tanımaktadır. Bu kaynakların bir kısmı, sadece otomatik oluşturulan tivitler atan yazılımlardır. Bu yüzden gerçek kullanıcılar tarafından oluşturulan tivitleri yakalamak için bazı kaynaklar seçilmiş ve diğer kaynaklardan gelen tivitler değerlendirme dışı bırakılmıştır. Sistemin topladığı tivitlere bakıldığında toplam 3.265 farklı kaynaktan geldikleri görülmektedir. Tivit Seçim Modülü, sadece şu kaynaklardan gelen tivitleri bir sonraki modüle aktarır:

- Twitter for Android
- Twitter Web Client

- Twitter for iPhone
- Twitter for iPad
- Twitter for Windows
- Twitter for Windows Phone
- Twitter for Android Tablets
- Mobile Web (M2)
- Mobile Web (M5)
- Facebook

Tivit seçim işleminin ikinci aşaması tivitleri kullanıcılar bazında değerlendirir. Burada amaç Gündem listesine bir etiket sokmak amacıyla aynı konuda, çok sayıda tivit atan kullanıcıların yakalanabilmesidir. Bunun için öncelikle tivitler kullanıcı bazında gruplanır. Daha sonra (3.1)'de gösterilen hesaplama yapılır.

$$O_i = T_{Pi} / T_{Ti} \quad (3.1)$$

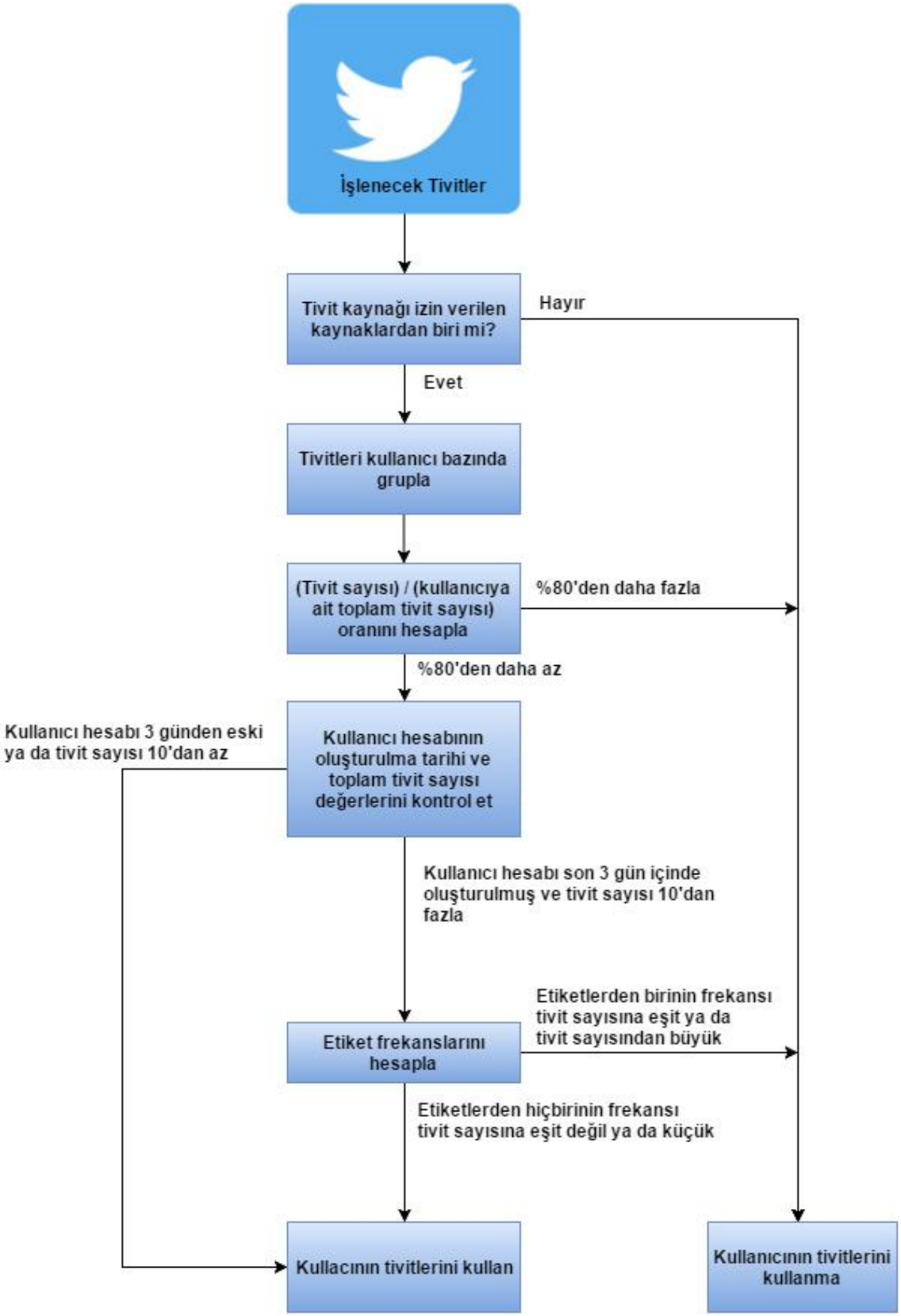
- O_i: i kullanıcısının program hakkında attığı tivit sayısının, kullanıcının toplam tivit sayısına oranı
- T_{Pi}: işlem yapılan yayın günü ve program için toplanan tivitlerden i kullanıcısına ait olan tivitlerin sayısı
- T_{Ti}: i kullanıcısının toplam tivit sayısı

O_i değeri hesaplandıktan sonra bu değerin %80 olarak belirlenen eşik değerinden büyük olup olmadığı kontrol edilir. Eğer bu değer eşik değerinden büyükse; yani kullanıcının bu program hakkında attığı tivitlerin sayısı, toplamda attığı tivit sayısının %80'inden fazlaysa, bu kullanıcının gündem oluşturma amaçlı, sürekli aynı konuda tivit atan sahte bir hesap olduğuna karar verilir ve kullanıcı tivitleri değerlendirme dışı bırakılır.

Üçüncü aşamada kullanıcı hesabının oluşturulma tarihine bakılır. Eğer hesap son 3 gün içinde oluşturulmuş ve işlem yapılan program hakkında 10 adetten fazla tivit paylaşmışsa, bu hesabın tek bir etiket hakkında tivitler paylaşmak için açılmış

olabileceği ihtimali değerlendirilir. Kullanıcılar, sevdikleri televizyon programları hakkındaki etiketleri Gündem listesine sokmak amacıyla, yayın günü çok sayıda hesap açarak aynı tivitleri paylaşmaktadırlar. Bunu kontrol etmek amacıyla tivitlerin içindeki etiketlerin frekansları hesaplanır. Eğer tivitlerin içindeki herhangi bir etiketin frekansı, tivit sayısına eşit ya da fazlaysa bu hesabın sadece bu etiket için açıldığına karar verilir ve tivitler değerlendirme dışı bırakılır.

Bütün bu kontrollerden geçebilen tivitler ise bir sonraki kısım olan Doğal Dil İşleme Modülü'ne aktarılır. Tivit Seçim Modülü tarafından yapılan işler, özet olarak Şekil 3.3'te gösterilmiştir.



Şekil 3. 3 Tivit seçim işlemi

3.4 Doğal Dil İşleme Modülü

Twitter'dan alınan veriler ham halleriyle oldukça kirli durumdadır. Sosyal medya jargonu, tivit içindeki linkler, 140 karakter sınırı nedeniyle yapılan kısaltmalar, yazım yanlışları gibi nedenlerden dolayı tivitlerin istem tarafından etiketlenmeden önce bazı işlemlerden geçirilmeleri gerekmektedir. Doğal Dil İşleme Modülü, tivitlerin Makine Öğrenmesi Modülü'ne gönderilmeden önce tek tek işlendiği ve temizlendiği yerdir.

Tivitler birkaç aşamada işlenmektedir. Öncelikle, tivitin elde edildiği sorgu için kullanılan arama kelimeleri ve durma sözcükleri (stop words) tivit metninden çıkarılmaktadır. Durma sözcükleri, cümlenin anlamını değiştirmeyen kelimelerdir (ile, ve, bazı, vb.). Daha sonra tivit içindeki bazı kalıplar bulunup, anahtar kelimeler ile değiştirilir. Bu kalıplar ve anahtar kelimeler şu şekildedir:

- Tivit içindeki linkler _URL_ anahtar kelimesi ile değiştirilir. Örneğin:

  İşlemden önce:

  "Evim Şahane yeni bölümüyle başlıyor! #evimsahane http://t.co/FUTa23Ywl5"

  İşlemden sonra:

  "Evim Şahane yeni bölümüyle başlıyor! #evimsahane _URL_"

- Tivit içindeki etiketler _HASHTAG_ anahtar kelimesi ile değiştirilir. Örneğin:

  İşlemden önce:

  "*Yaşasınnn #banaherseyyakısır başlamışşşşş:)"*

  İşlemden sonra:

  "*Yaşasınnn _HASHTAG_ başlamışşşşş:)"*

- Tivit içindeki tarihler _DATE_ anahtar kelimesi ile değiştirilir. Örneğin:

  İşlemden *önce:*

  "25 Şubat 2015 Çarşamba Fox Esra Erol'la http://t.co/ukcb7sZ1Gq"

  İşlemden sonra:

  "_DATE_ Fox Esra Erol'la http://t.co/ukcb7sZ1Gq"

- Tivit içindeki kullanıcı adları _USERNAME_ anahtar kelimesi ile değiştirilir. Örneğin:

  İşlemden önce:

  "*@fatihportakal fox haber basliyorrrrr*"

  İşlemden sonra:

  "*_USERNAME_ fox haber basliyorrrrr*"

- Tivit içindeki noktalama işareti ile oluşturulan ifadeler ve duygu ifade eden ikonlar ("☺", "☹", ":)", "😍", ":(" vb.) anlamlarına göre _POSWORD_ (olumlu anlamlılar) ya da _NEGWORD_ (olumsuz anlamlı) anahtar kelimeleri ile değiştirilir. Bu işlem için birer olumlu ve olumsuz ifadeler listesi hazırlanmıştır. Sistem, bir tivit içinde bu ifadelere rastladığında anlamına göre uygun anahtar kelime ile değiştirir. Örneğin:

  İşlemden önce:

  "bu tarz benim izliyorum 😒 nur yerlitaş ♥"

  İşlemden sonra:

  "bu tarz benim izliyorum _NEGWORD_ nur yerlitaş _POSWORD_"

Kalıpları anahtar kelimelerle değiştirme işlemi bittikten sonra tivit metni içindeki noktalama işaretleri silinir ve kelimeler için Türkçeleştirme (deasciify) işlemi yapılır. Twitter kullanıcılarının çoğu mobil olarak siteye erişmekte ve tivit paylaşmaktadır [14]. Klavyesinde Türkçe'de kullanılan bazı harfleri (ı, İ, ç, ş, ö, ü) barındırmayan cep telefonlarından tivit paylaşan kullanıcılar kelimeleri bu harfler olmadan yazmak zorunda kalmaktadırlar ("ışık" yerine "isik" yazılması vb.). Doğal Dil İşleme Modülü, bu kelimeleri Türkçeleştirme işlemi ile düzeltmektedir. Örneğin:

İşlemden önce:

"Fox'ta gunahkar yeni bolum... cok guzel dizi"

İşlemden sonra:

"Fox'ta günahkar yeni bölüm... çok güzel dizi"

Bir sonraki işlem yazım yanlışlarının düzeltilmeye çalışılmasıdır. Bunun için Zemberek [16] kütüphanesi kullanılmıştır. Zemberek kütüphanesi içindeki kelime denetleme fonksiyonu, verilen herhangi bir kelimenin Türkçe bir kelime olup olmadığını denetleyebilmektedir. Yazım yanlışı içeren kelimeler bu fonksiyon tarafından geçerli Türkçe kelimeler olarak tanınmamaktadır. Bu yüzden bu fonksiyon kullanılarak tivitler içindeki (anahtar kelimeler hariç) kelimeler tek tek denetlenmiş ve yanlış kelimeler ayrılmıştır. Zemberek kütüphanesi, bir öneri fonksiyonu da barındırmaktadır. Bu fonksiyon yanlış yazılan bir kelime yerine, uzaklık algoritması kullanarak, kelimenin doğrusu olabilecek Türkçe kelimeler önermektedir. Kelime denetleme fonksiyonu tarafından tanınmayan kelimeler öneri fonksiyonuna verilerek, doğru halleri elde edilmeye çalışılmıştır. Öneri fonksiyonu tek bir öneride bulunduysa, kelime yerine bu öneri kullanılmaktadır. Fakat fonksiyon geriye birden fazla öneri de döndürebilmektedir. Bu öneriler algoritma tarafından en yakın kelimeden en uzağa doğru sıralanmış halde getirilmektedir. Örneğin “akşm” kelimesi öneri fonksiyonuna verildiğinde dönen sonuçlar Çizelge 3.1’de gösterilmiştir.

Çizelge 1. 1 “akşm” kelimesi için Zemberek tarafından yapılan öneriler

Öneri No	Öneri
1	akşam
2	akım
3	akam
4	aküm
5	AKM
6	AKM'm

Bu çalışmada, birden fazla öneri varsa ilk gelen öneri hatalı yazılan kelime yerine kullanılarak ilerlenmiştir. Eğer öneri fonksiyonu tarafından herhangi bir sonuç döndürülmezse kelime içinde tekrar eden harf olup olmadığına bakılmıştır. Tivitlerin içinde bu tür kelimelere sıkça rastlanmaktadır. Bazen anlamı kuvvetlendirmek için (canım yerine canımmmm yazmak gibi), bazen de yazım yanlışları nedeniyle bu kelimeler var olmaktadır. Kelime içindeki tekrar eden harfler tek adede indirilip, öneri fonksiyonuna tekrar gönderilmekte ve tekrar aynı işlemler yapılmaktadır. Fonksiyondan yine sonuç dönmediği durumda kelime olduğu haliyle bırakılmıştır.

Son işlem olarak kelimelerin kökleri alınmıştır. Bütün kelimeler yerine sadece kökler ile işlem yapıldığında daha yüksek sınıflandırma başarısı elde edilmektedir. Kökleri bulabilmek için tekrar Zemberek Kütüphanesi kullanılmıştır. Kütüphane içindeki kök bulma ve kelime çözümleme fonksiyonları ile kelimeye ait kök ve ekler elde edilmiştir. Örneğin "izlemiyorum" kelimesi için bu fonksiyonlar kullanıldığında elde edilen sonuçlar Çizelge 3.2'de gösterilmiştir:

Çizelge 2. 2 "izlemiyorum" kelimesi için çözümleme sonucu

Değer	**Tür**
izle	FIIL SDK_ZAMAN
mi	FIIL_OLUMSUZLUK_ME
yor	FIIL_SIMDIKIZAMAN_IYOR
um	FIIL_KISI_BEN

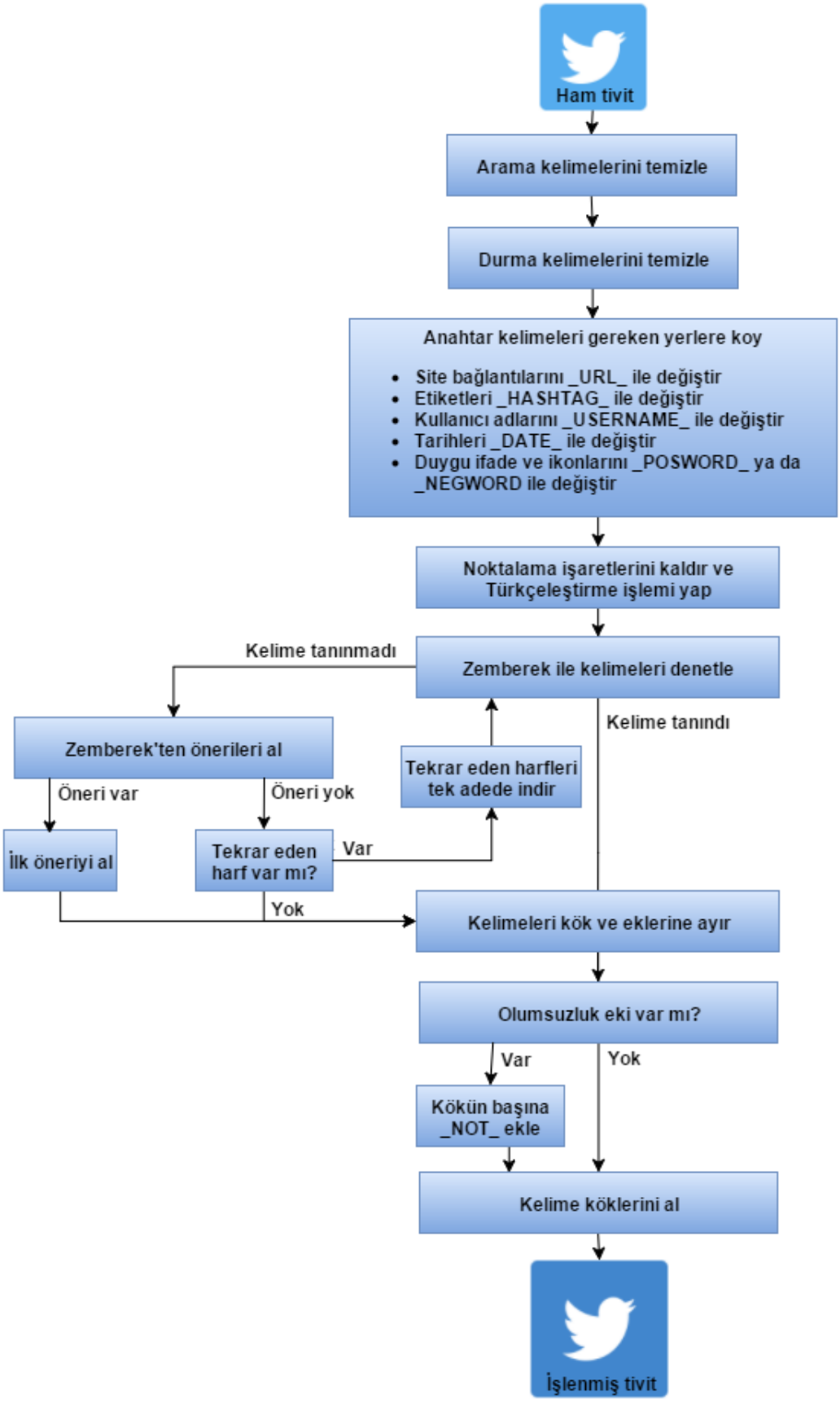
Kelime çözümlendikten sonra elde edilen kök, kelime yerine kullanılmaktadır. Fakat Türkçe sondan eklemeli bir dil olduğundan, bir kelimenin olumlu ya da olumsuz kullanılmış hallerinin kökleri aynı olmaktadır. Bu durumu aşmak amacıyla elde edilen ekler arasından olumsuzluk eki olup olmadığına bakılmaktadır. Eğer kelime içinde olumsuzluk eki mevcutsa kök, başına _NOT_ anahtar kelimesi eklenerek kullanılmaktadır. Çizelge 3.3'te "izliyorum" ve "izlemiyorum" kelimeleri için elde edilen kökler örnek olarak gösterilmiştir.

Çizelge 3. 3 "izliyorum" ve "izlemiyorum" kelimeleri için elde edilen kök değerleri

Kelime	**Kök Değeri**
izliyorum	izle
izlemiyorum	_NOT_izle

Yukarıda anlatılan işlemler, Tivit Seçim Modülü tarafından Doğal Dil İşleme Modülü'ne aktarılan tüm tivitler için uygulanmaktadır. İşlenen tivitler Makine Öğrenmesi Modülü'ne aktarılır.

Doğal Dil İşleme Modülü tarafından yapılan işlemler özet olarak Şekil 3.4'teki gibidir.



Şekil 3. 4 Doğal Dil İşleme Modülü işlem akışı

3.5 Makine Öğrenmesi Modülü

Makine Öğrenmesi Modülü, Doğal Dil İşleme Modülü'nden gelen tivitlerin sınıflandırılarak etiketlendiği bölümdür. Sınıflandırma işleminde, veri madenciliği için kullanılan bir makine öğrenmesi algoritmaları kütüphanesi olan Weka [17] aracından faydalanılmıştır. Weka, işlem yapabilmek için verilerin "ARFF" (Attribute-Relation File Format) isminde özel bir formatta olmasına ihtiyaç duymaktadır. Bu yüzden sınıflandırılması için modüle verilen veriler ARFF formatına dönüştürülerek kullanılmaktadır.

Bu çalışmada Weka içindeki 3 farklı algoritma kullanılmıştır. Bunlar:

- Naive Bayes
- Destek Vektör Makinesi
- Rastgele Orman algoritmalarıdır.

Modül, seçilen algoritmayı daha önceden oluşturulan bir eğitim seti ile eğitir. Eğitim seti oluşturma ile ilgili detaylar Bölüm 4.2'de verilecektir. Daha sonra verilen herhangi bir tivit, algoritma tarafından 3 sınıftan (olumlu, olumsuz ve nötr) biri olarak sınıflandırılır ve sınıflandırma sonucu kaydedilir. Sistemdeki tüm tivitler için bu sınıflandırma işlemi yapılır.

3.5.1 Sınıflandırma Algoritmaları

Bu bölümde, Makine Öğrenmesi Modülü tarafından kullanılan sınıflandırma algoritmaları hakkında bilgiler verilecektir. Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi algoritmalarının çalışma mantıkları anlatılacaktır.

3.5.1.1 Naive Bayes (NB)

Naive Bayes, metin sınıflandırma işlemlerinde sıklıkla kullanılan ve hızlı çalışan bir yöntemdir. Bayes teoremini temel alan bu yöntem bir veri örneğinin (X), var olan n adet sınıftan (S_1, S_2, S_3, …, S_n) hangisine ait olduğunu belirlemek için kullanılır. Her veri örneği çok boyutlu özellik vektörleri ile gösterilmektedir ($X = X_1, X_2, X_3, \ldots, X_k$). Her özellik eşit

öneme sahiptir ve birbirinden bağımsızdır. Özellikler, birbirleri hakkında bilgi barındırmazlar. Bayes teoremi (3.2)'deki gibi ifade edilmektedir.

$$P(Ci|X)=\frac{P(X|Ci)P(Ci)}{P(X)} \quad (3.2)$$

Yukarıdaki denklemde gösterilen değerler şunlardır,

- $P(C_i|X)$ değeri X örneğinin C_i sınıfından olması olasılığı
- $P(X)$ değeri X örneğinin olasılığı
- $P(C_i)$ değeri C_i sınıfının olasılığı
- $P(X|C_i)$ değeri C_i sınıfından bir örneğin x örneği olma olasılığı

P(X) değerinin tüm sınıflar için sabit olduğu durumda X örneğinin C_i sınıfından olma olasılığı (3.3) ile hesaplanabilir.

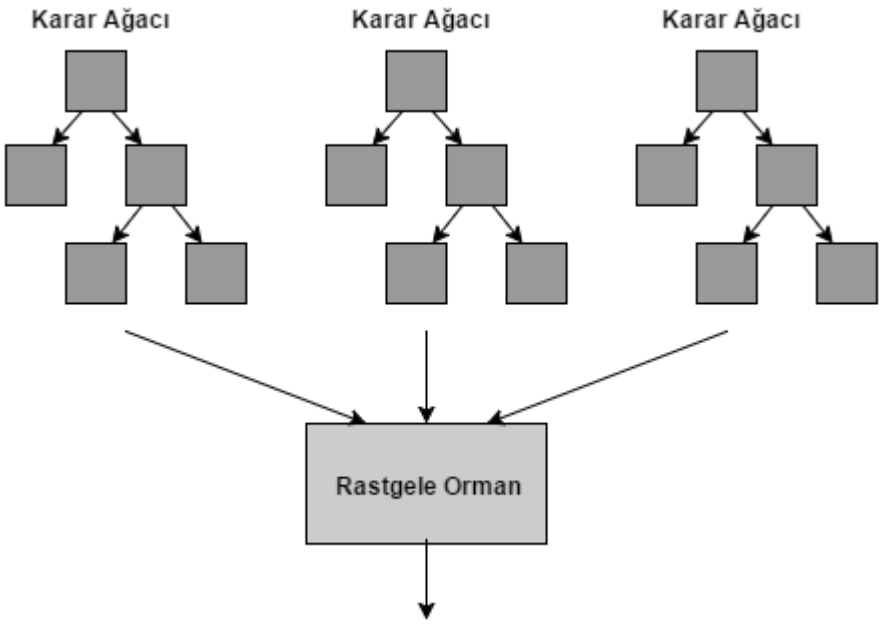
$$P(C_i|X) = P(X|C_i)P(C_i) \quad (3.3)$$

NB yönteminde, tüm sınıflar için $P(C_i|X)$ değerini hesaplanmaktadır. Daha sonra en yüksek sonucu veren sınıf, X örneğinin bulunduğu sınıf olarak seçilir. Bu işlem (3.4)'deki şekilde ifade edilebilir.

$$P(X|C_i) = \prod_{k=1}^{m} P(X_k|C_i) \quad (3.4)$$

3.5.1.2 Rastgele Orman (RO)

Rastgele Orman [18], Breiman ve Cutler tarafından sunulmuş ve birden fazla karar ağacını bünyesinde barındıran bir sınıflandırma algoritmasıdır. Diğer toplu sınıflandırma yöntemlerinde olduğu gibi birden fazla sınıflandırıcı ile yapılan tahmin sonuçlarını kullanarak sınıflandırma sonucu vermektedir. RO, tahmin yapmak için en popüler sonucu baz alır. Ağaçlarda düğümler dallara ayrılırken mevcut özelliklerin hepsi değil, rastgele seçilen bir alt kümesi içindeki en iyi sonucu verecek olan özellik kullanılarak işlem yapılır. Şekil 3.5'te RO sınıflandırıcısının ağaç yapısı verilmiştir.



Şekil 3. 5 Rastgele Orman ağaç yapısı

Bu çalışmada kullanılan RO yönteminde 10 adet ağaç oluşturularak ilerlenmiştir.

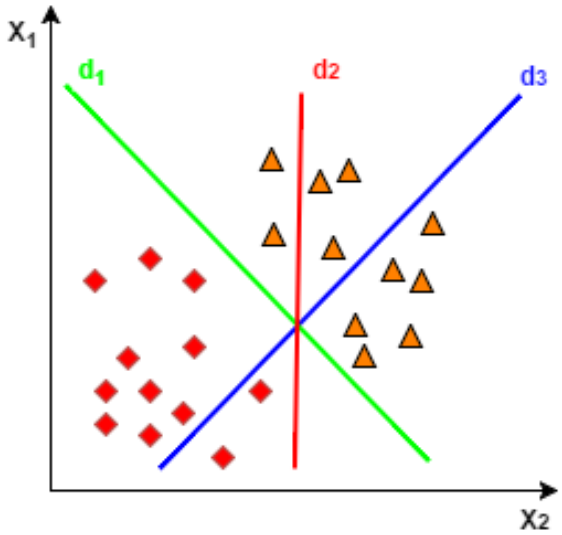
3.5.1.3 Destek Vektör Makinesi (DVM)

Destek Vektör Makinesi, çalışmada sınıflandırma için kullanılan üçüncü yöntemdir. Bu yöntem, sınıflandırma, aykırı değer bulma ve regresyon için kullanılabilmektedir. Yapısal risk minimizasyonu ve istatistiksel öğrenme teorisine dayanan bu yöntem, veri içindeki değişkenler arası örüntülerin bilinmediği veri setlerinde sınıflandırma için kullanılabilmektedir. Yöntemin sahip olduğu avantajlar şunlardır:

- Karmaşık karar sınırlarını modelleyebilmektedir.
- Çok sayıda bağımsız değişkenle çalışabilmektedir.
- Doğrusal olarak ayrılan verilere uygulanabildiği gibi doğrusal olarak ayrılamayan verilere de uygulanabilmektedir.
- Aşırı uyum sorunu, diğer birçok yönteme göre daha azdır.

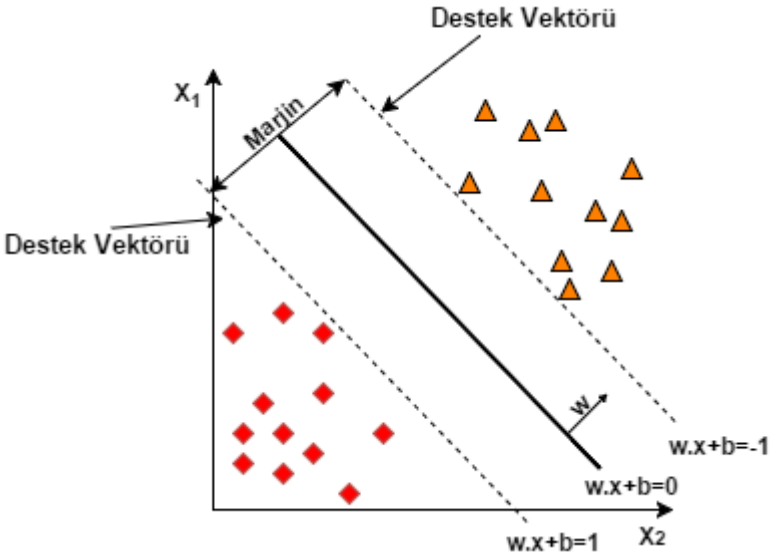
Doğrusal olarak ayrılabilen ve doğrusal olmayan olarak iki gruba ayırabileceğimiz DVM yöntemi temel olarak iki sınıflı problemlerle ilgilenmektedir. Doğrusal olarak ayrılabilen DVM yönteminde amaç bir veri setini iki sınıfa ayırabilecek olan sonsuz sayıda doğrudan en uygununu seçmektir. Şekil 3.6'da hiperdüzlem adı verilen bu doğrulara örnek olarak

d1, d2 ve d3 doğruları gösterilmiştir. Burada, d2 ve d3 doğruları başarılı bir ayırma işlemi gerçekleştiremezken, d1 doğrusu iki veri sınıfını başarıyla ayırmıştır.



Şekil 3. 6 DVM'de verileri ayrıştıran doğrular

DVM'de sınıfları birbirinden aynı başarıyla ayırabilen hiperdüzlemlerden en yüksek marjine sahip olan hiperdüzlem seçilmektedir. En yüksek marjine sahip olan hiperdüzlemi bulabilmek için her iki sınıfın en uygun doğruya en yakın veri noktalarından geçen ve birbirine paralel olan doğrular çizilir. Bu doğrulara destek vektörleri adı verilir. Destek vektörleri arasındaki uzaklığa da marjin denmektedir. Şekil 3.7, bu kavramları göstermektedir.



Şekil 3. 7 DVM tarafından çizilen destek vektörleri, marjin ve hiperdüzlem

İki sınıfa ait verileri birbirinden ayıran hiperdüzlem için w.x+b=0 ifadesi doğru olmaktadır.

Destek vektörler ise (3.5)'teki gibi ifade edilmektedir.

$w.x + b = 1$ (3.5)

$w.x + b+ = -1$

Marjin genişliği, yani destek vektör hiperdüzlemleri arasındaki uzaklık, 2/w olarak ifade değerindedir. Bu yüzden w değeri ne kadar küçülüşe marjin genişliği o kadar artmaktadır. DVM yönteminde amaç en yüksek marjin genişliğini bulmak olduğu için |w| değeri minimize edilmeye çalışılmaktadır. Veri noktaları x_i ve sınıf değerleri y (-1, +1) kullanılarak eğitilen bir sistemin en uygun hiperdüzlemi (3.6)'daki şekilde bulunur.

$y = +1 \text{ ise } w.x_i + b >= 1$ (3.6)

$y = -1 \text{ ise } w.x_i + b < 1$

Buradaki tüm i değerleri için $y_i(w.x_i + b) >= 1$ denilebilmektedir. Sınıflandırma işleminde herhangi bir x örneği için, bulunan hiperdüzlemin hangi tarafına düştüğüne göre etiketleme işlemi yapılır. Hiperdüzlemin bir tarafı pozitif, bir tarafı ise negatif sınıf olarak adlandırılır. Eğer x örneği pozitif tarafta ise pozitif sınıfa, negatif tarafta ise negatif sınıfa ait olarak etiketlenir.

Bu çalışmada kullanılan veriler, 2 değil 3 farklı sınıftan oluşmaktadır. Bu yüzden iki sınıf için çalışan DVM ile sınıflandırma yapılırken birden fazla sınıflandırıcı oluşturularak ilerlenmiştir. Sistemdeki 3 farklı sınıf (olumlu, olumsuz ve nötr) için, pozitif/negatif, pozitif/nötr ve negatif/nötr sınıfları için çalışan 3 farklı sınıflandırıcı oluşturulmuştur. Böylece çoklu sınıf yerine ikili sınıflarla işlem yapılmış, sınıflandırma sonucu olarak bu 3 sınıflandırıcıdan alınan sonuçların birleşimi kullanılmıştır.

3.6 Sıralama Tahmin Modülü

Sıralama Tahmin Modülü, yayınlanan programlar için günlük bir izlenme oranı listesi oluşturan modüldür. Bu modül, Makine Öğrenmesi Modülü tarafından sınıflandırılan tivit bilgilerini kullanarak yayınlanan her program için gün bazında bir puan

hesaplamaktadır. Daha sonra programlar, bu puanlara göre sıralanarak o gün için izlenme oranı tahmini yapılmış olur.

Hesaplama işlemi, bir yayın günündeki her program için şu veriler kullanılarak yapılmaktadır:

- Olumlu tivit sayısı
- Olumsuz tivit sayısı
- Toplam tivit sayısı
- Olumlu tivitlerin retweet edilme sayısı
- Olumlu tivitlerin favorilenme sayısı
- Olumsuz tivitlerin retweet edilme sayısı
- Olumsuz tivitlerin retweet edilme sayısı
- Olumlu ve olumsuz tivit oranı (PTO – NTO) katsayısı
- Toplam Tivit Oranı (TTO) katsayısı

Puan hesaplama işlemi birkaç adımda yapılmaktadır. Öncelikle, Makine Öğrenmesi Modülü tarafından olumlu, olumsuz ya da nötr olarak etiketlenmiş olan tivitlerin sayısı gün bazında alınmaktadır. Bu aşamada tivitlerin favori ve retweet sayıları da dikkate alınmaktadır. Bir tivitin başka bir kullanıcı tarafından retweet edilmesi ya da favorilenmesi, kullanıcının o tivitte anlatılana katıldığını göstermektedir. Bu yüzden, tivitlerin retweet ve favori sayıları tivitin sınıfına göre toplam olumlu, olumsuz ya da nötr tivit sayısına eklenmektedir. Daha sonra program için TTO (Toplam Tivit Oranı puanı) hesaplanır. (3.7), bu hesaplamayı göstermektedir.

$$S_{Ti} = (T_{MAX} / T_{i)} * F_T \quad (3.7)$$

Yukarıdaki denklemde gösterilen değerler şunlardır:

- S_{Ti}: Puanı hesaplanacak olan programa ait günlük TTO değeri
- T_i: Programın günlük toplam tivit sayısı

- T_{MAX}: Hesaplama yapılan gün içinde hakkında en fazla tivit atılmış olan programın toplam tivit sayısı
- F_T: TTO katsayısı

TTO hesaplanırken T_{MAX} değerinin kullanılmasının sebebi programlar hakkında atılan tivit sayılarının çok farklı olabilmesidir. Örneğin bazı günler, izlenme oranlarında ilk sırada olan bir programın toplam tivit sayısı rakiplerinden çok daha az da olabilmektedir. Yani bir program hakkında çok sayıda tivit atılması o programın çok izlediğini göstermeyebilmektedir. Bu yüzden, TTO hesaplanırken, programların toplam tivit sayıları, hesap yapılan gün hakkında en çok tivit atılmış olan programın tivit sayısına oranlanmaktadır. Böylece bir programın, sadece çok sayıda tivit sahibi olduğu için diğerlerinin çok önüne geçmesi önlenmektedir.

TTO hesaplandıktan sonra PTO (Pozitif Tivit Oranı puanı) ve NTO (Negatif tivit oranı puanı) değerleri hesaplanmaktadır. PTO değerinin hesaplanması (3.8)'de gösterilmiştir.

$$S_{Pi} = ((T_{Pi} + R_{Pi} + F_{Pi}) / T_i) * F_{PN} \quad (3.8)$$

Yukarıdaki denklemde gösterilen değerler şunlardır:

- S_{Pi}: Puanı hesaplanacak olan programa ait günlük PTO değeri
- T_{Pi}: Programın günlük toplam pozitif tivit sayısı
- R_{Pi}: Programa ait pozitif tivitlerin retweet sayısı
- F_{Pi}: Programa ait pozitif tivitlerin favorilenme sayısı
- T_i: Programın günlük toplam tivit sayısı
- F_{PN}: PTO - NTO katsayısı

NTO değerinin hesaplanması (3.9)'da gösterilmiştir.

$$S_{Ni} = ((T_{Ni} + R_{Ni} + F_{Ni}) / T_i) * F_{PN} \quad (3.9)$$

Yukarıdaki denklemde gösterilen değerler şunlardır:

- S_{Ni}: Puanı hesaplanacak olan programa ait günlük NTO değeri
- T_{Ni}: Programın günlük toplam negatif tivit sayısı

- R_{Ni}: Programa ait negatif tivitlerin retweet sayısı
- F_{Ni}: Programa ait negatif tivitlerin favorilenme sayısı
- T_i: Programın günlük toplam tivit sayısı
- F_{PN}: PTO - NTO katsayısı

Bu işlemlerden sonra program için günlük puan hesaplaması (3.10)'daki ki şekilde yapılmaktadır.

$$S_i = S_{Ti} + S_{Pi} - S_{Ni} \quad (3.10)$$

Yukarıdaki denklemde gösterilen değerler şunlardır:

- S_i: GPO (Günlük Program Puanı)
- S_{Ti}: Programın günlük TTO değeri
- S_{Pi}: Programın günlük PTO değeri
- S_{Ni}: Programın günlük NTO değeri

Yukarıdaki hesaplamalarda kullanılan F_T ve F_{PN} değerleri, GPO'da toplam tivit sayısına göre (TTO) ve anlama göre (PTO - NTO) hesaplanan alt puanların ağırlıklarını belirlemek için kullanılan katsayılardır. Bu katsayıların kullanılma sebebi program hakkında atılan tivit sayısı ile atılan pozitif ya da negatif tivit sayısının izlenme oranı üzerindeki etkisinin aynı olmamasıdır. Bu katsayıların kullanılmadığı durumda, çok tivit sayısına sahip olan programlar çok puan almaktadır. Örneğin, hakkında %70'i negatif, %30'u pozitif olmak üzere toplam 20.000 tivit atılmış olan bir program, hakkında %70' i pozitif olan 5.000 tivit atılmış bir programdan daha az izlenmiş olabilmektedir. Program hakkında atılan pozitif tivit sayısının önemi, toplam tivit sayısından büyük olabilmektedir. Bu katsayı değerlerinin seçimi Bölüm 5.4.2'de anlatılmaktadır.

Sıralama Tahmin Modülü, yukarıda bahsedilen hesaplamaları, o gün yayınlanan tüm programlar için yaptıktan sonra elde edilen GPO değerlerine göre azalan bir sıralama oluşturmakta ve kaydetmektedir.

3.7 Sonuç Görüntüleme Modülü

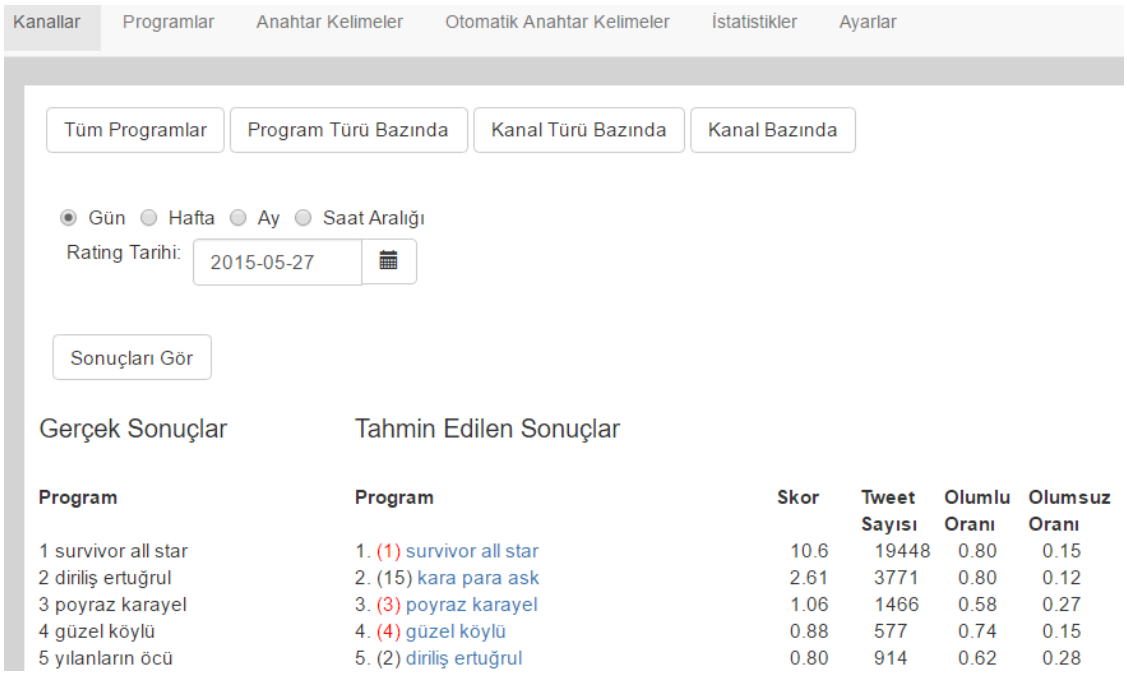
Sonuç Görüntüleme Modülü, kullanıcının sistem ile iletişime geçtiği ara yüzü barındıran modüldür. Bu modül ASP.NET Web API [19] ve AngularJS [20] teknolojileri kullanılarak geliştirilmiş bir web uygulamasıdır. Modül üzerinden şu bilgiler görüntülenebilmektedir:

- Tahmini izlenme oranı sıralamaları
- Gerçek izlenme oranı sıralamaları
- Program bazında hesaplanan puanlar
- Program bazında toplam tivit sayıları, olumlu ve olumsuz tivit oranları
- Sistem içindeki kanallar, programlar, arama terimleri ve detayları

Sistemin hesapladığı tahmini izlenme oranı sıralamaları şu kriterler ile görüntülenebilmektedir:

- Kanal
- Kanal türü (Genel, haber, müzik, spor vb.)
- Program türü (Dizi, yarışma, haber, belgesel, magazin, müzik vb.)
- Gün, hafta, ay
- Yayın kuşağı (Gündüz kuşağı, gece kuşağı, prime time vb.)
- Verilen herhangi bir zaman aralığı

Modülün web ara yüzü üzerinden sisteme yeni kanal, program ve arama terimi eklenebilmekte, var olanlar güncellenebilmekte ya da silinebilmektedir. Şekil 3.8, sonuç görüntüleme ekranını göstermektedir.



Şekil 3. 8 Sonuç Görüntüleme Modülü sonuç görüntüleme ekranı

BÖLÜM 4

EĞİTİM SETLERİ

Bu bölümde, Bölüm 3.5'te bahsedilen Makine Öğrenmesi Modülü tarafından kullanılacak olan en iyi sınıflandırma yöntemini seçebilmek amacıyla yapılacak deneylerde kullanılmak üzere oluşturulan eğitim setlerinden bahsedilmiştir. 4.1'de eğitim setlerini oluşturmak için kullanılacak olan etiketli veri setinin hazırlanması; 4.2'de eğitim setlerinin oluşturulması anlatılmaktadır.

4.1 Etiketli Veri Seti Oluşturulması

Eğitim setlerindeki verilerin sınıf etiketine sahip veriler olması gerekmektedir. Etiketli bir veri setinin oluşturulması için öncelikle bölüm 3.1'de detayları verilmiş olan Veri Toplama Modülü kullanılarak toplanan tivitler içerisinden, rastgele bir alt küme seçilmiştir. Seçilen bu alt kümedeki tivitler, toplandıkları halleriyle okunup, şu 3 sınıftan biri olarak etiketlenmiştir:

- Olumlu
- Olumsuz
- Nötr

Tivitlere verilecek olan etiketler kişisel fikirlere göre farklılık gösterebileceğinden dolayı etiketleme işlemi 3 farklı kişi tarafından gerçekleştirilmiştir. Etiketlenen bu tivitlerden her sınıf için 1.000'er adet farklı tivit alınarak toplamda 3.000 adetlik, tekrarsız bir veri seti oluşturulmuştur. Bu veri seti çeşitli işlemlerden geçirilerek 4 farklı eğitim seti elde edilmiştir.

4.2 Eğitim Setlerinin Oluşturulması

İlk eğitim seti, etiketli veri setinin, detayları Bölüm 3.4'te verilen Doğal Dil İşleme Modülü içindeki bütün işlemlerden geçirilmesi ile elde edilen özelliklerden oluşturulmuştur. Bu eğitim setine "Sözcük Eğitim Seti" adı verilmiştir.

Diğer eğitim setlerini oluşturmak için n-gram yöntemi kullanılmıştır. N-gram, bir dizi karakterin ardışık sıralı n adetlik elemanlarıdır. Bu yöntemle veri içindeki örüntüleri bulmak amaçlanmaktadır. N değeri kaç alınırsa o kadar karakterlik alt kümeler oluşturulur. Örneğin "ana haber" ifadesinin 2-gram ve 3-gram değerleri şu şekilde olmaktadır:

2-gram'lar: *"an", "na", "a_", "_h", "ha", "ab", "be", "er"*

3-gram'lar: *"ana", "na_", "a_h", "_ha", "hab", "abe", "ber"*

N-gram yöntemini uygulamadan önce etiketli veri seti Doğal Dil İşleme Modülü tarafından işlemden geçirilmiştir. Fakat bu işlemde modülün kök alma özelliği kullanılmamıştır. Böylece, kelimeler üzerinde düzeltme işlemleri yapılmış ve örüntü kaybına neden olmamak için kelimeler bütün halde tutulmuştur. Daha sonra, n değeri 2 ve 3 olarak alınarak elde edilen özellikler ile "2gram Eğitim Seti" ve "3gram Eğitim Seti" adında eğitim setleri oluşturulmuştur. En son, bu iki eğitim setindeki özellikler birleştirilerek "2gram + 3gram Eğitim Seti" isminde dördüncü bir eğitim seti daha oluşturulmuştur.

Oluşturulan 4 eğitim seti, Makine Öğrenmesi Modülü tarafından kullanılan Weka kütüphanesinin tanıyabileceği format olan ARFF formatına dönüştürülerek kaydedilmiştir.

Eğitim setlerindeki özellik sayıları birbirinden farklı olmuştur. Çizelge 4.1, liste bazında özellik sayılarını göstermektedir.

Çizelge 4. 1 Eğitim setlerinin özellik sayıları

Eğitim Seti Adı	**Özellik Sayısı**
Sözcük Eğitim Seti	3713
2gram Eğitim Seti	1107
3gram Eğitim Seti	5957
2gram + 3gram Eğitim Seti	7064

BÖLÜM 5

DENEYSEL SONUÇLAR

Bu bölümde geliştirilen sistem ve sistem ile yapılan uygulamaların deneysel sonuçları verilmiştir. Bölüm 5.1'de, sistem içindeki veriler ile ilgili istatistikler verilmektedir. Bölüm 5.2'de, veri setleri üzerinde test edilen özellik azaltma yöntemi ve bu yöntem kullanılarak yapılan test sonuçlarından bahsedilmektedir. Bölüm 5.3'te sınıflandırma algoritmaları ile yapılan testlerin karşılaştırmalı sonuçları; Bölüm 5.4'te ise sistemin tahmin başarısına ilişkin sonuçlar verilmektedir. Bölüm 5.2 ve 5.3'te verilen başarı değerleri "F-ölçüm" yöntemi kullanılarak hesaplanmıştır. F-ölçüm, ikili sınıflandırıcıların istatistiksel analizinde doğruluğu göstermek için kullanılan, kesinlik (precision) ve hassasiyet (recall) değerleri kullanılarak hesaplanan bir değerdir.

5.1 Sistem Verileri

Sistem, birçok farklı türde veri barındırmakta ve bu verileri kullanarak çalışmaktadır. Sistemdeki veriler şunlardan oluşmaktadır:

- Tivitler
- Twitter kullanıcıları
- Kanallar
- Programlar
- Arama terimleri
- Gerçek izlenme oranı puanları ve sıralamaları
- Tanımlar (Kanal, program, arama terimi türleri ve geçerli tivit kaynakları)

Bu verilerden tivitler, Twitter kullanıcıları ve gerçek izlenme oranları Veri Toplama Modülü tarafından 01-11-2014 / 28-06-2015 tarihleri arasında internetten toplanmıştır. Diğer veri türlerine ait veriler ise sisteme kullanıcı tarafından elle kaydedilmiş ya da sistem tarafından otomatik olarak oluşturulmuş verilerdir. Sistemdeki verilerin türlere göre adetleri Çizelge 5.1'de gösterilmektedir:

Çizelge 5. 1 Sistemdeki verilerin türlere göre adetleri

Veri Türü	**Adet**
Tivit	6.310.257
Twitter kullanıcısı	1.124.540
Kanal	15
Program	839
Arama terimi	3.351
Kanal türü	4
Program türü	20
Arama terimi türü	5
Tivit kaynağı	3.265
Gerçek program izlenme oranı	16.636

Sistem tarafından toplanan tivitlerin 3.265 adet farklı tivit kaynağından geldiği görülmüştür. Bölüm 3.3.1'de bahsedildiği şekilde, sadece 10 adet tivit kaynağı sistem tarafından geçerli kaynaklar olarak işaretlenmiş ve sadece bu kaynaklardan gelen tivitler hesaplama işlemlerinde kullanılmaktadır. Sistem tarafından toplanan toplam 6.310.257 adet tivitten, 5.228.318 tanesinin bu geçerli kaynaklardan geldiği görülmüştür. Bu 5.228.318 adet tivitin ise 1.484.986 tanesi Tivit Seçim Modülü tarafından elenmiştir. Kalan 3.743.332 adet tivit hesaplama işlemlerinde kullanılmıştır. Böylece sistemin hesaplamalar için kullandığı tivitlerin, sistem tarafından toplanan toplam tivit sayısına oranı yaklaşık olarak %59,3 olmuştur.

5.2 Özellik Azaltma

Bölüm 4.2'de, nasıl oluşturuldukları anlatılan eğitim setleri içindeki özelliklerden hepsi verileri sınıflandırmada ayırt edici olmamaktadır. Bu ayırt edici olmayan özelliklerin kullanılması ile özellik uzayının büyümesi, sınıflandırma başarısını olumsuz etkileyebilmektedir. Sadece ayırt edici olan özelliklerin bulunup kullanılması amacıyla

eğitim setleri üzerinde özellik azaltma işlemi uygulanarak daha az özellik barındıran veri setleri elde edilmiştir. Özellik azaltma yöntemi olarak Weka içindeki CfsSubsetEval [21] fonksiyonu kullanılmıştır.

Özellik azaltma yöntemi uygulandıktan sonra elde edilen veri setlerindeki özellik sayıları büyük oranda azalmıştır. Çizelge 5.2, veri setlerinin özellik azaltma uygulanmış ve uygulanmamış durumlarındaki özellik sayılarını ve özellik sayılarındaki yaklaşık azalma oranlarını vermektedir.

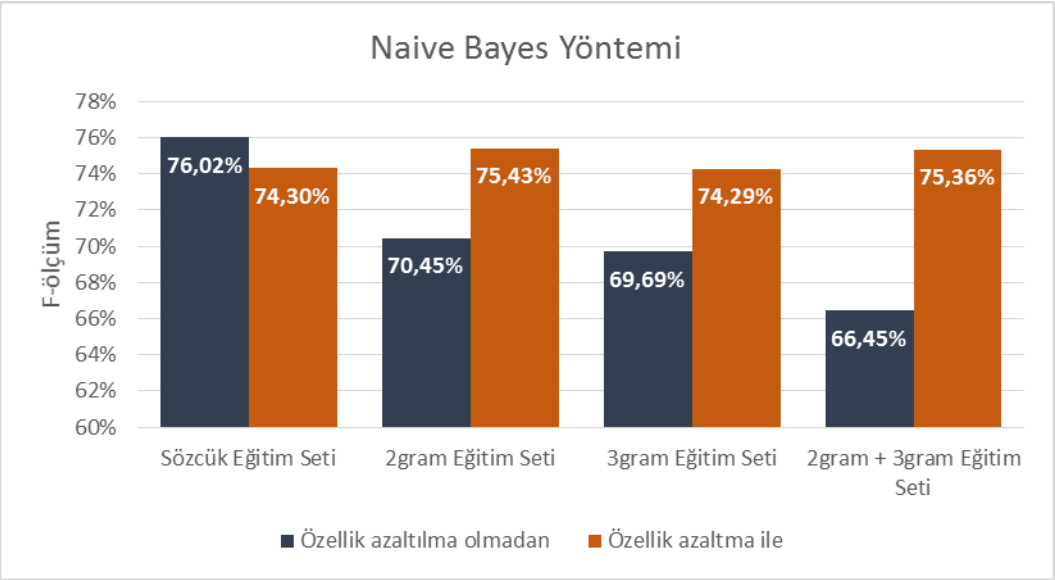
Çizelge 5. 2 Veri setlerinin özellik azaltma uygulanmadan önce ve sonraki özellik sayıları ve oranları

Veri Seti Adı	**Tüm Özelliklerin Sayısı**	**Azaltılmış Özellik Sayısı**	**Özellik Azalma Oranı**
Sözcük Eğitim Seti	3713	52	%98,59
2gram Eğitim Seti	1107	34	%96,92
3gram Eğitim Seti	5957	51	%99,14
2gram + 3gram Eğitim Seti	7064	54	%99,24

Sistemdeki 3 sınıflandırıcı için özellik azaltma uygulanmış ve uygulanmamış veri setleri ile 10 katlı çapraz doğrulama yapılarak, doğru sınıflandırma oranları alınmıştır.

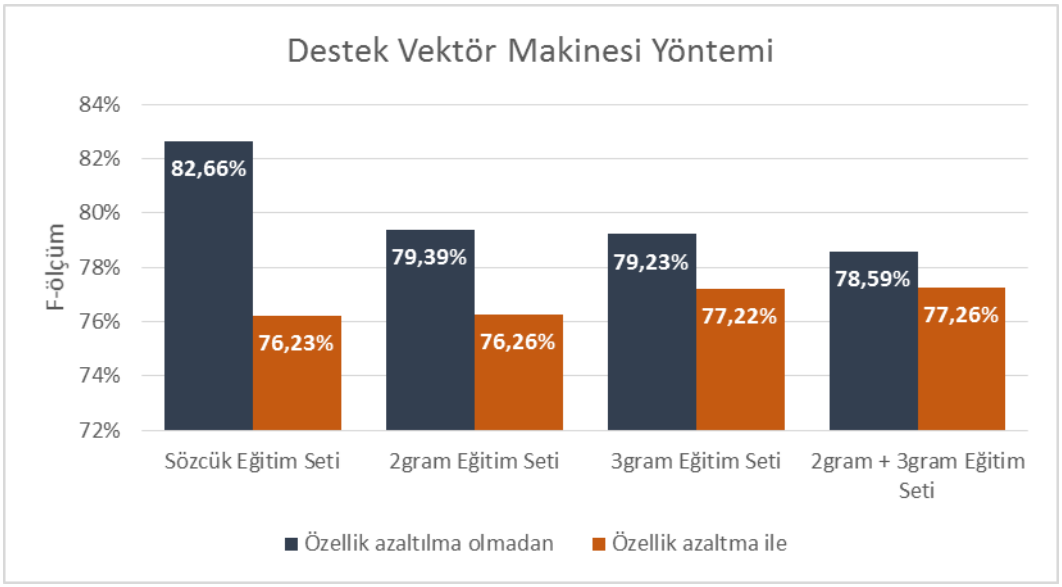
N katlı çapraz doğrulama yöntemi, eğitim verisini n adet gruba böler. N defa tekrar eden bir süreç ile her grup bir kez test verisi olarak kullanılır. Sınıflandırıcı, diğer n-1 adet grup ile eğitilir ve test verisi olarak ayrılan grup ile test edilerek başarısı hesaplanır. Sonuç, elde edilen n adet başarı hesabının ortalaması olarak alınır. Bu çalışmada yapılan testlerde n değeri 10 olarak alınmıştır.

Şekil 5.1, Naive Bayes yönteminin, özellik azaltma ve uygulanmış ve uygulanmamış veri setleri ile elde ettiği doğru sınıflandırma oranlarını göstermektedir. Özellik azaltma işleminin, Sözcük Eğitim Seti kullanıldığında sınıflandırma başarısını %1,72 oranında düşürdüğü gözlemlenmiştir. Fakat diğer veri setleri üzerinde özellik azaltma yöntemi başarıyı arttırmıştır. En yüksek artış, %8.91'lik oranla 2gram + 3gram Eğitim Seti üzerinde meydana gelmiştir.



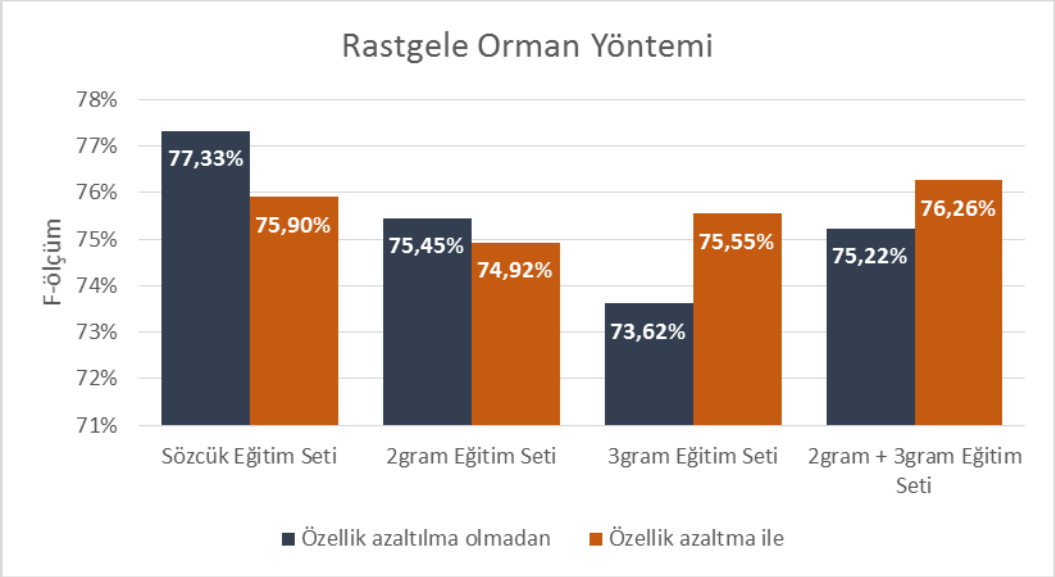
Şekil 5. 1 Özellik azaltma kullanımının farklı veri setleri ile eğitilen Naive Bayes sınıflandırıcısının başarısı üzerindeki etkisi

Destek Vektör Makinesi kullanılarak yapılan testlerde özellik azaltma işlemi, tüm veri setleri için başarıyı azaltıcı bir etki yapmıştır. Şekil 5.2, Destek Vektör Makinesi yönteminin, özellik azaltma uygulanmış ve uygulanmamış veri setleri ile elde ettiği doğru sınıflandırma oranlarını göstermektedir. Özellik azaltma işleminin bütün veri setleri için başarıyı düşürdüğü görülmektedir. En büyük düşüş %6,43 ile Sözcük Eğitim Seti isimli veri seti kullanıldığında meydana gelmiştir.



Şekil 5. 2 Özellik azaltma kullanımının farklı veri setleri ile eğitilen Destek Vektör Makinesi sınıflandırıcısının başarısı üzerindeki etkisi

Şekil 5.3'da Rastgele Orman yönteminin, özellik azaltma uygulanmış ve uygulanmamış veri setleri ile elde ettiği doğru sınıflandırma oranlarını göstermektedir. Özellik azaltma işleminin diğer yöntemlerde olduğu gibi Sözcük Eğitim Seti kullanıldığında başarıyı düşürdüğü görülmektedir. 2gram Eğitim Seti kullanıldığında da özellik azaltma işleminin başarı üzerinde olumsuz etkisi olmuştur. 3gram Eğitim Seti ve 2gram + 3gram Eğitim Seti isimli veri setlerinde ise özellik azaltma başarıyı arttırmıştır.



Şekil 5. 3 Özellik azaltma kullanımının farklı veri setleri ile eğitilen Rastgele Orman sınıflandırıcısının başarısı üzerindeki etkisi

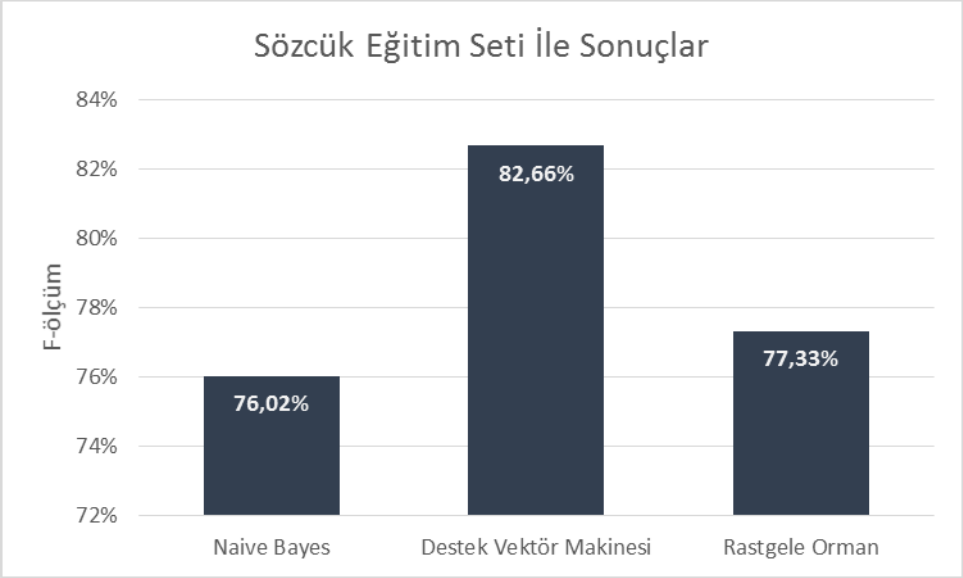
Bu deneyler sonucunda özellik seçiminin başarıya olan etkisinin kullanılan sınıflandırma algoritması ve veri setine göre değiştiği gözlemlenmiştir. 3 sınıflandırıcı için de ortak olan nokta, özellik seçiminin Sözcük Eğitim Seti kullanıldığında başarı üzerindeki olumsuz etkisidir.

5.3 Sınıflandırıcı Karşılaştırılması ve Seçimi

Sınıflandırma başarılarını ölçmek için sistemde kullanılan 3 farklı sınıflandırıcı ile deneyler yapılmıştır. Bu deneylerde sınıflandırıcılar, Bölüm 4.2'de oluşturulma aşamaları anlatılan Sözcük Eğitim Seti, 2gram Eğitim Seti, 3gram Eğitim Seti ve 2gram + 3gram Eğitim Seti adlı veri setleri ile eğitilmiştir. Bölüm 5.3'te sonuçlarından bahsedilen özellik azaltma işlemi ise başarıyı arttırdığı sınıflandırıcı ve veri seti eşleştirmelerinde kullanılmış, başarıyı düşürdüğü durumlarda kullanılmamıştır. Daha sonra 10 katlı çapraz doğrulama yapılarak başarı sonuçları elde edilmiştir. Bu deneylerin amacı en yüksek

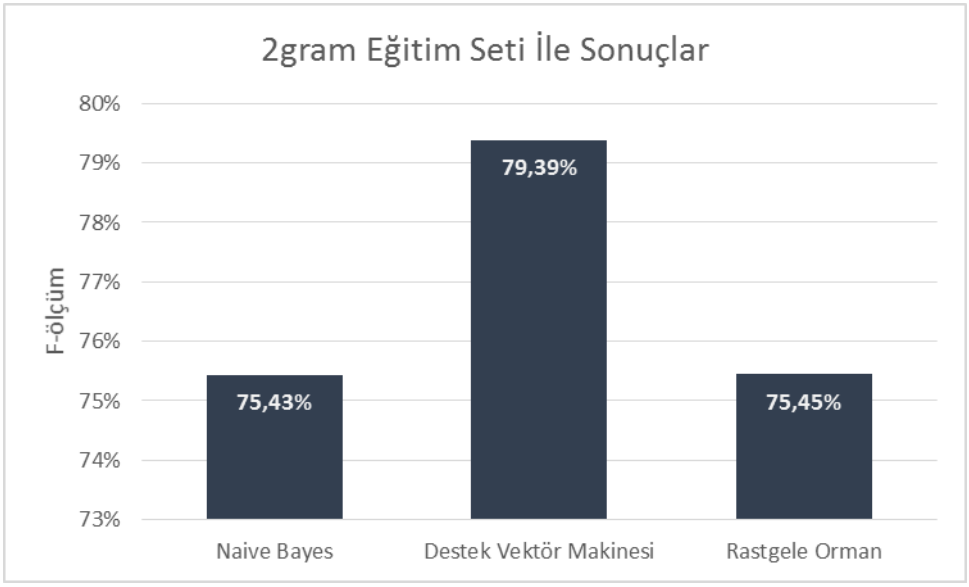
başarıyı veren sınıflandırıcı ve veri seti yöntemini bulup, Bölüm 3.5'te anlatılan Makine Öğrenmesi Modülü tarafından bu yöntemlerin kullanılmasını sağlamak olmuştur.

Şekil 5.4'te Naive Bayes, Destek Vektör Makinesi ve Rastgele Orman sınıflandırıcıları Sözcük Eğitim Seti kullanılarak test edildiğinde elde edilen başarılar gösterilmektedir. Sınıflandırıcılar, veri seti üzerinde özellik azaltma işlemi yapılmadan test edilmiştir. Bu veri seti için en başarılı sonucu %82,66 doğru sınıflandırma oranı ile Destek Vektör Makinesi almıştır. En düşük başarı ise %76,02 ile Naive Bayes sınıflandırıcısının olmuştur. Rastgele Orman yöntemi ise Naive Bayes yönteminden biraz daha yüksek başarı göstermesine karşın Destek Vektör Makinesi yönteminin elde ettiği başarının oldukça gerisinde kalmıştır.



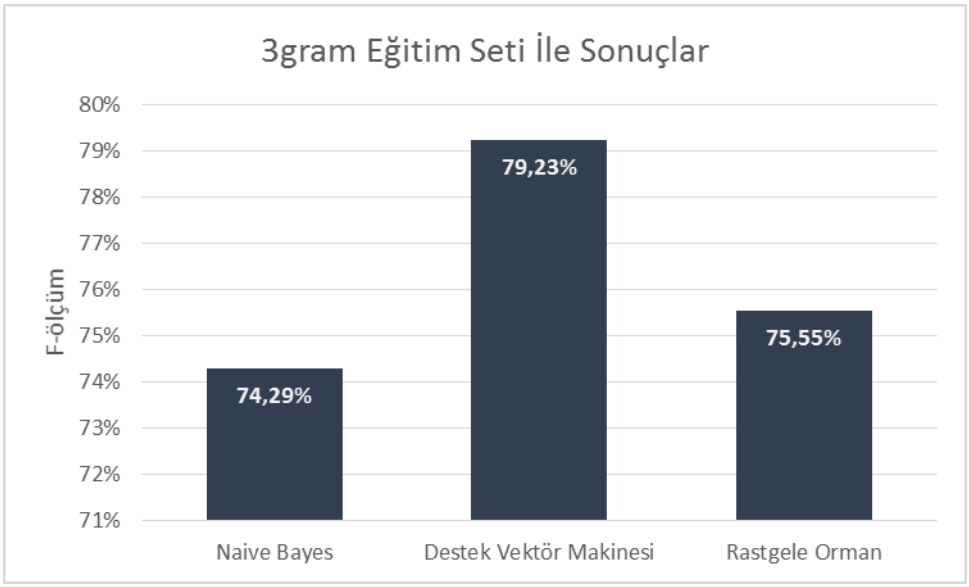
Şekil 5. 4 Sözcük Eğitim Seti kullanılarak test edilen sınıflandırıcıların başarıları

Şekil 5.5'te Naive Bayes, Destek Vektör Makinesi ve Rastgele Orman sınıflandırıcıları 2gram Eğitim Seti kullanılarak test edildiğinde elde edilen başarılar gösterilmiştir. Bu deneyde Naive Bayes sınıflandırıcısı özellik azaltma işleminden geçirilen veri seti kullanılarak test edilmiştir. Diğer 2 sınıflandırıcı tarafından kullanılan veri seti için ise özellik azaltma işlemi uygulanmamıştır. Bu deneyde de en yüksek başarı, %79,39 doğru sınıflandırma oranı ile Destek Vektör Makinesi yöntemi tarafından elde edilmiştir. En düşük başarı %75,43 ile Naive Bayes tarafından gösterilmiştir. Rastgele Orman yönteminin başarısı %75,45 ile çok Naive Bayes'e çok yakındır.



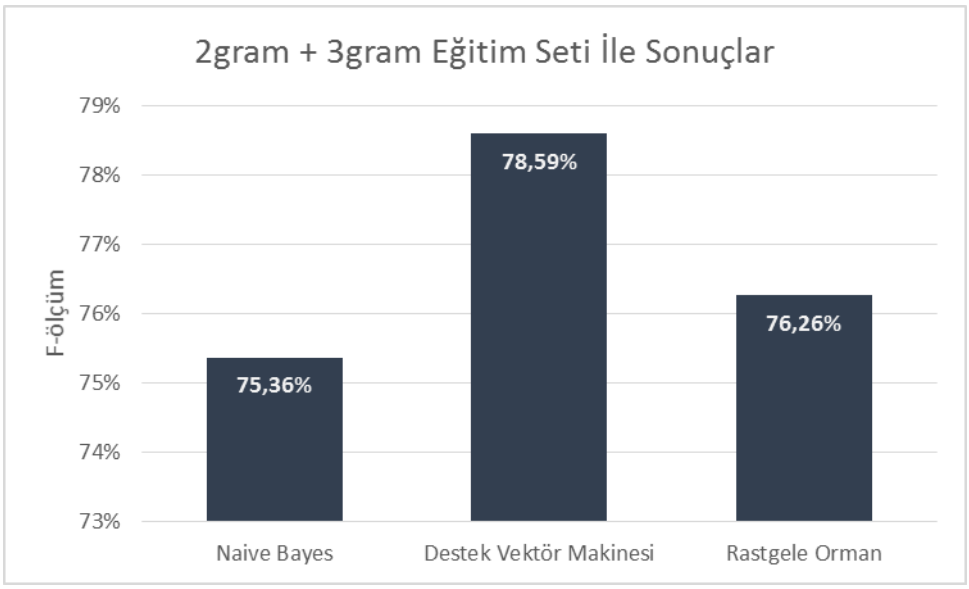
Şekil 5. 5 2gram Eğitim Seti kullanılarak test edilen sınıflandırıcıların başarıları

Şekil 5.6'da Naive Bayes, Destek Vektör Makinesi ve Rastgele Orman sınıflandırıcıları 3gram Eğitim Seti kullanılarak test edildiğinde elde edilen başarılar gösterilmiştir. Naive Bayes ve Rastgele orman sınıflandırıcıları test edilirken özellik azaltma işleminden geçirilen veri seti kullanılmıştır. Bu deneyde de en yüksek başarı, %79,23 doğru sınıflandırma oranı ile Destek Vektör Makinesi'ne aittir. En düşük başarı %74,29 ile Naive Bayes tarafından gösterilmiştir. Rastgele Orman yöntemi ise %75,55 başarı ile ikinci sıradadır.



Şekil 5. 6 3gram Eğitim Seti kullanılarak test edilen sınıflandırıcıların başarıları

Şekil 5.7'de Naive Bayes, Destek Vektör Makinesi ve Rastgele Orman sınıflandırıcıları 2gram + 3gram Eğitim Seti kullanılarak test edildiğinde elde edilen başarılar gösterilmiştir. Bu deneyde de, başarıyı arttırdığından dolayı, Naive Bayes ve Rastgele orman sınıflandırıcıları test edilirken özellik azaltma işleminden geçirilen veri seti kullanılmıştır. En yüksek başarı yine Destek Vektör Makinesi tarafından %78,59 doğru sınıflandırma oranı ile elde edilmiştir. En düşük başarı %75,36 ile Naive Bayes tarafından gösterilmiştir. Rastgele Orman yöntemi ise %76,26 başarı ile ikinci sıradadır.



Şekil 5. 7 2gram + 3gram Eğitim Seti kullanılarak test edilen sınıflandırıcıların başarıları

Bazılarında özellik azaltma işlemi uygulanan, 4 farklı veri seti ve 3 farklı sınıflandırıcı kullanılarak yapılan deney başarılarına bakılarak, tüm deneylerde en başarılı sınıflandırıcının Destek Vektör Makinesi olduğu görülmüştür. En yüksek başarı elde edilen veri seti ise her deneyde Sözcük Eğitim Seti olmuştur. Tüm deneylerde en düşük başarıyı gösteren sınıflandırıcı Naive Bayes olarak gözlemlenmiştir. Rastgele Orman, genel olarak Naive Bayes'in biraz üstünde başarı gösterse de Destek Vektör Makinesi'nin oldukça gerisinde kalmıştır. Özellik azaltma işlemi bazı sınıflandırıcı ve veri setleri için başarıyı arttırmış, bazıları için düşmüştür.

Sonuç olarak en yüksek başarı %82,66 ile özellik azaltma uygulanmamış Sözcük Eğitim Seti isimli veri setini kullanan Destek Vektör Makinesi sınıflandırıcısı ile elde edilmiştir. Bu bilgiye dayanarak geliştirilen sistem kullanılırken, Bölüm 3.5'te detaylandırılan Makine Öğrenmesi Modülü'nde sınıflandırıcı olarak Destek Vektör Makinesi yöntemi

kullanılmıştır. Sistemin çalışması esnasında sınıflandırılması gereken tüm tivitler, özellik azaltma işlemi uygulanmamış Sözcük Eğitim Seti ile eğitilmiş olan bu sınıflandırıcı kullanılarak sınıflandırılmıştır.

5.4 İzlenme Oranı Sırası Tahmin Başarısı

Bu bölümde sistemin izlenme oranı sıralamaların tahmin edebilme başarısını ölçmek için yapılan çalışmalar ve sonuçlar anlatılmaktadır. Sırasıyla başarıyı ölçmek için kullanılan yöntemler tanımlanacak, izlenme oranı puanı hesaplamasında kullanılan katsayıların seçimi için yapılan deneyler anlatılacak ve izlenme oranı sırası tahmin sonuçları paylaşılacaktır.

5.4.1 İzlenme Oranı Sırası Tahmin Başarısı Ölçme Yöntemleri

Sistemin, izlenme oranı sıralaması tahmin başarılarını ölçmek için iki farklı yöntem kullanılmıştır. Bu yöntemlerden ilki "İlk-n" ismini verdiğimiz yöntemdir. Bu yöntemde, her yayın günü için resmi sonuçlarda izlenme oranı listesinin ilk n sırasındaki programın, sıralama bağımsız olarak doğru tahmin edilme başarısını ölçmektir. Bir yayın günü için resmi sonuçlarda ilk n sırada olan programlardan, tahmin edilen listede ilk n sırada olanların oranı hesaplanır. Veri toplanmış olan tüm günler için bu hesaplama yapıldıktan sonra ortalama alınarak başarı elde edilir.

Sistemin tahmin başarısını ölçmek için kullanılan ikinci yöntem, "Ortalama Fark Puanı" adını verdiğimiz ve her programın tahmin edilen sıralamasının, resmi sonuçtaki sıralamasına göre farkına dayanan bir puan hesaplamasıdır. Örneğin; tahmin listesinde 3. sırada olan bir program, resmi listede de 3. sırada ise bu program için hesaplanan puan 1 olur. Ama bu program resmi sıralamada 4. sırada ise, yani 1 sıralık bir fark varsa, puan 0,5 olarak hesaplanır. Tüm programlar için puan hesabı yapıldıktan sonra ortalama puan alınarak Ortalama Fark Puanı hesaplanmış olur. (5.1), tek bir yayın günü için puan hesaplamasını göstermektedir:

$$P = \sum_{i=0}^{n} \frac{1}{|Sti - Sri| + 1} \tag{5.1}$$

Yukarıdaki denklemde gösterilen değerler şunlardır,

- P: Bir yayın günü için alınan fark puanı
- n: Tahmin edilen izlenme oranı sıralamasındaki program sayısı
- Sti: i. programın tahmin edilen izlenme oranı listesindeki sırası
- Sri: i. programın resmi izlenme oranı listesindeki sırası

Bu yöntemde alınan her bir programın aldığı puan, resmi listedeki ile tahmin edilen listedeki sıralaması arasındaki fark büyüdükçe azalmaktadır. Böylece tahmin ne kadar doğruysa o kadar yüksek puan alınması sağlanmıştır. Tüm program sıralamaları doğru tahmin edilmiş olan bir yayın günü için elde edilen puan 1 olacaktır.

5.4.2 TTO ve PTO - NTO Katsayısı Seçimleri

Bölüm 3.6'da yapılan hesaplamalarda bazı katsayılar kullanılmıştır. Hesaplanan izlenme oranı puanı içinde toplam tivit oranı ve anlama göre sınıflandırılmış (pozitif ya da negatif) tivitlerin oranının ağırlığını belirleyen bu katsayıları seçmek için bazı testler yapılmıştır.

En başarılı katsayıları bulmak için Bölüm 5.4.1'de bahsedilen Ortalama Fark Puanı yöntemi kullanılmıştır. Test için öncelikle her iki katsayı için de 1'den 10'a kadar tüm tamsayı değerleri kullanılarak Bölüm 3.6'da bahsedilen yöntemlerle izlenme oranı puanları hesaplanmıştır. Daha sonra bu puanlara göre oluşturulan izlenme oranı listelerindeki sıralamalar ile resmi izlenme oranı sonuçlarındaki sıralamalar için Ortalama Fark Puanı yöntemi kullanılarak başarı puanları hesaplanmıştır.

Çizelge 5.3, test edilen katsayı değerlerinden en yüksek başarıyı elde eden ilk 5 katsayı ikilisini ve bu katsayıların Ortalama Fark Puanı değerlerini göstermektedir.

Çizelge 5. 3 En yüksek başarıyı gösteren 5 katsayı ikilisi ve Ortalama Fark Puanı değerleri

TTO Katsayısı	PTO – NTO Katsayısı	Ortalama Fark Puanı
10	1	0,1444
9	1	0,1440
8	1	0,1436
7	1	0,1423
6	1	0,1420

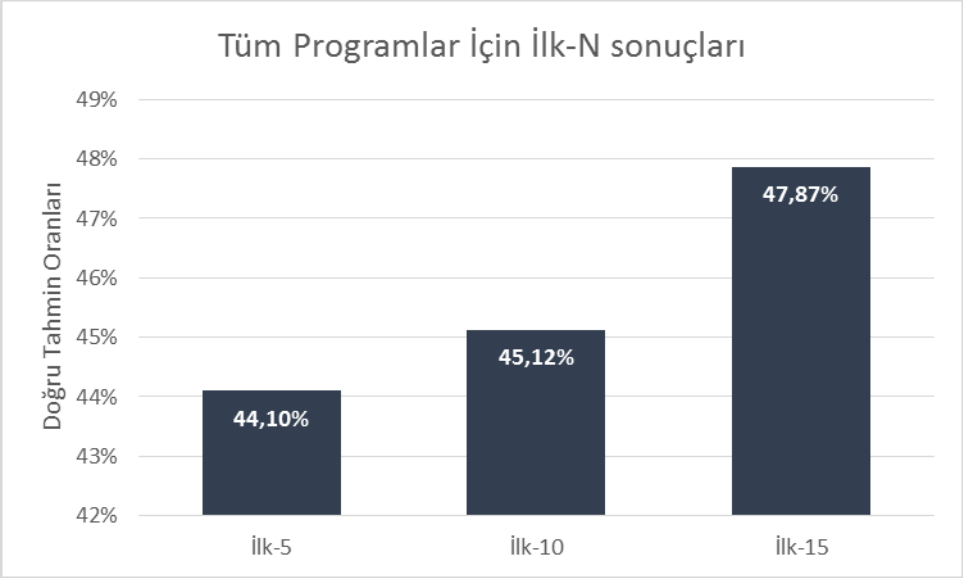
Elde edilen sonuçlara göre TTO katsayısı 10, PTO- NTO katsayısı 1 iken Ortalama Fark Puanı 0,1444 olmuştur. Bu değer bu test için elde edilmiş olan en yüksek değerdir. Bu sonuca dayanarak sistemin tahmin başarısı ölçülürken bu katsayı değerleri kullanılmıştır.

5.4.3 İzlenme Oranı Sırası Tahmin Sonuçları

Bu bölümde sistem tarafından yapılan izlenme oranı sıralama sonuçları ile resmi izlenme oranı sıralamaları karşılaştırılmış ve ölçülen tahmin başarıları paylaşılmıştır. Bölüm 3.6'da bahsedilen Sıralama Tahmin Modülü ile hesaplanan izlenme oranı puanlarına göre oluşturulan sıralamalar, sistem tarafından verilerin toplandığı 01-11-2014 / 28-06-2015 tarihleri arasına aittir. Bu bölümde paylaşılan sonuçlar, bu zaman aralığını kapsamaktadır.

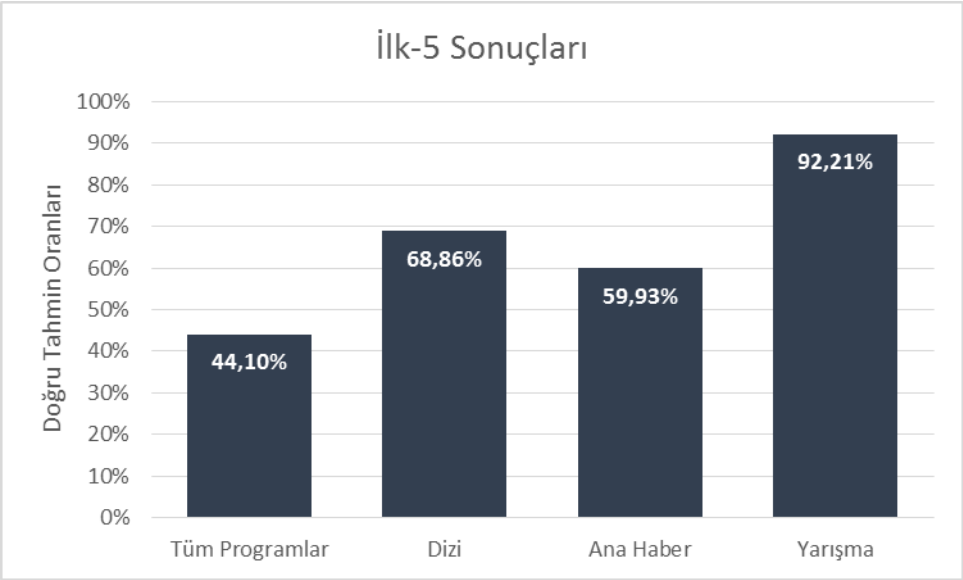
Sistemin tahmin başarısını ölçmek için Bölüm 5.4.1'de bahsedilen İlk-n ve Ortalama Fark Puanı yöntemleri kullanılmıştır.

Şekil 5.8, sistemdeki tüm programlar için İlk-5, İlk-10 ve İlk-15 test yöntemi ile elde edilen sonuçları göstermektedir. Bu sonuçlara göre n sayısı arttıkça tahmin ile resmi sonuçların uyuşma oranının arttığı görülmektedir. İlk-5 yöntemi ile test yapıldığında ilk 5 sıradaki programların ortalama olarak %44,10'unun doğru tahmin edildiği görülmektedir. İlk-10 yöntemi için %45,12 olan bu oran, İlk-15 yöntemi için ise %47,87 ile en yüksek halini almıştır.



Şekil 5. 8 Tüm programlar için İlk-5, İlk-10 ve İlk-15 yöntemi ile yapılan testlerde elde edilen sonuçlar

Tüm programlar için yapılan testlerin dışında, farklı program türleri için de testler yapılmıştır. Şekil 5.9, dizi, ana haber ve yarışma türündeki programlar için de ilk 5 tahmin başarı sonuçlarını göstermektedir. Dizi ve yarışma türündeki programlar için sonuçlar her gün, ana haber türündeki programlar için sonuçlar ise sadece, ana haber bültenlerinin yayınlandığı günler olan, hafta içleri dikkate alınarak hesaplanmıştır. Bu sonuçlara göre program türü bağımsız olarak test yapıldığında ilk 5 sıradaki programların ortalama olarak %44,10'unun doğru tahmin edildiği görülmüştür. Dizi türündeki programlar için ilk 5 sıradaki programlar %68,86 ortalama ile tahmin edilebilmiştir. Ana haber türünde ise başarı oranı %59,93 olmuştur. Yarışma türündeki programlar için ise %92,21'lik oranla oldukça yüksek bir başarı elde edilmiştir. Bu verilere dayanarak, program türünün sistem tahmin başarısı üzerinde büyük etkisi olduğu söylenebilmektedir. Bunun sebebinin, program türlerine göre paylaşılan tivit sayılarındaki değişiklikler olduğu düşünülmektedir. Özellikle yarışma türü için elde edilen yüksek başarının sebebinin yarışma programları hakkında diğer türlerden daha çok tivit paylaşılması olduğu söylenebilir.



Şekil 5. 9 Tüm programlar ve dizi, ana haber, yarışma türleri için İlk-5 yöntemi ile yapılan testlerde elde edilen sonuçlar

Sistemin tahmin başarısını ölçmek için kullanılan ikinci yöntem Ortalama Fark Puanı yöntemidir.

Çizelge 5.4'te, yine ilk 5 sıradaki programlar dikkate alınarak dizi, ana haber, yarışma, magazin ve spor programları için ikinci yöntem kullanılarak hesaplanan puanlar gösterilmektedir. Ana haber türündeki programların puanı, yine sadece yayınlandığı günler olan hafta içi günleri dikkate alınarak hesaplanmıştır.

Çizelge 5. 4 Farklı program türleri için hesaplanan başarı puanları

Program Türü	**Ortalama Fark Puanı**
Dizi	0,3694
Ana Haber	0,3757
Yarışma	0,6283

Elde edilen sonuçlara bakıldığında ana haber türündeki programların dizi türündeki programlara göre daha yüksek bir puan elde ettiği görülmektedir. İlk-5 yöntemine göre elde edilen sonuçlarda ise dizi türü için ana haber türüne göre daha yüksek başarı elde edildiği ve iki yöntemin sonuçlarının bu iki program türü için uyumsuz olduğu görülmektedir. Fakat yarışma türündeki programlar için elde edilen Ortalama Fark Puanı, İlk-5 testinde olduğu gibi dizi ve ana haber programlarının puanlarından çok daha fazla olarak, tutarlılık göstermiştir.

BÖLÜM 6

SONUÇ VE ÖNERİLER

Bu çalışmada, en büyük sosyal medya veri kaynaklarından biri olan Twitter üzerinden elde edilen veriler kullanılarak Türk televizyonları izlenme oranı sıralaması tahmini yapabilen bir sistem geliştirilmeye çalışılmıştır.

Sistem tarafından 01-11-2014 / 28-06-2015 tarihleri arasında çeşitli kaynaklardan veri toplanmıştır. Toplanan tivitler oldukça kirli durumda olduğundan üzerlerinde çeşitli temizlik işlemleri yapılmıştır. Ayrıca bu tivitlerden sadece gerçekten kullanıcılar tarafından paylaşılmış ve onların fikirlerini yansıtan tivitlerin seçilmesi için bazı eleme işlemleri yapılmıştır. Bu eleme işlemleri sonucu toplanan 6.310.257 adet tivitten sadece 3.743.332 adet tivitin sistem tarafından kullanılması uygun görülmüştür.

Tivitler sınıflandırma işleminden geçirilmiş ve elde edilen veriler ile her program için izlenme oranı puanları hesaplanmıştır. Bu puanlar kullanılarak tahmini izlenme oranı listeleri oluşturulmuştur.

Sistemde kullanılacak olan sınıflandırma algoritmasını seçmek amacıyla Naive Bayes, Destek Vektör Makinesi ve Rastgele Orman sınıflandırıcıları ile çeşitli veri setleri kullanılarak testler yapılmıştır. Kullanılan veri setleri, tivitlerin n-gram yöntemi kullanılarak ve kullanılmadan işlenmiş halleri ile oluşturulmuştur. Ayrıca korelasyon tabanlı özellik seçiminin sınıflandırma başarısı üzerindeki etkileri incelenerek, en yüksek başarı elde edilmeye çalışılmıştır.

Yapılan testler sonucunda en yüksek sınıflandırma başarısı Destek Vektör Makinesi sınıflandırma algoritmasında, özellik azaltma işlemi uygulanmayan Sözcük Eğitim Seti isimli veri seti kullanılarak elde edilmiştir. Elde edilen %82,66'lık sınıflandırma başarısı

neticesinde geliştirilen sistem tarafından Destek Vektör Makinesi ana sınıflandırma algoritması olarak kullanılmıştır. N-gram yöntemi ile elde edilen veri setlerinde sınıflandırma başarısının, 3 sınıflandırma algoritması için de daha düşük olduğu gözlemlenmiştir. Özellik azaltma yönteminin ise bazı algoritmalar ve veri setleri için başarı üzerinde olumlu etkisi olduğu görülmüştür. Özellik azaltma işlemi, uygulandığı tüm veri setleri için özellik sayısında %95'ten daha büyük oranda azalma sağlamıştır.

İzlenme oranı tahmin sonuçlarının ölçülebilmesi için İlk-n ve Ortalama Fark Puanı isimli iki yöntem geliştirilmiştir.

Her yayın günü ve program için hesaplanan izlenme oranı puanı hesaplamasında kullanılan ve toplam tivit oranı ile anlamsal olarak sınıflandırılmış tivitlerin oranlarının bu puan içindeki ağırlıklarını belirlemeye yarayan TTO ve PTO – NTO katsayı değerlerinin seçilmesi için Ortalama Fark Puanı yöntemi ile testler yapılmıştır. Bu testler sonucunda en yüksek başarı elde edilen değerler olmaları sebebiyle sistem tarafından kullanılacak olan değerler, TTO katsayısı değeri için 10, PTO- NTO katsayısı değeri için 1 olarak seçilmiştir.

İzlenme oranı sıralama tahmini başarısını ölçmek için İlk-5 yöntemi kullanarak yapılan testlere göre program türünün sistemin tahmin başarısı üzerinde büyük bir etkisi olduğu söylenebilmektedir. Dizi türünde olan programlar kullanılarak elde edilen test sonucuna göre, ilk 5'e giren programların ortalama olarak %68,86'sı sistem tarafından tahmin edilebilmiştir. Yarışma türündeki programlar için yapılan İlk-5 testinde ise başarı oranı %92,21'e olmuştur. Bu durum, program türünün başarı üzerindeki etkisini göstermektedir. Yarışma türü için elde edilen yüksek başarının sebebinin bu tür programlar için diğer türlere göre daha fazla tivit paylaşılması olduğu düşünülmektedir. Program türü bağımsız olarak yapılan İlk-5 testi sonucuna göre ise programların %44,10'unun doğru tahmin edildiği görülmüştür.

Ortalama Fark Puanı yöntemi ile yapılan başarı ölçümlerinde ise dizi türündeki programlar için 0,3694 puan elde edilmiştir. Yarışma türündeki programlar için ise elde edilen puan 0,6283 olmuş ve bu iki program türü için İlk-5 testindeki sonuçlarla tutarlılık göstermiştir.

Sınıflandırma algoritmalarını eğitmek için kullanılan eğitim seti boyutu arttırılarak sistemin sınıflandırma başarısını arttırmak mümkün olabilir. Böylece daha doğru tahmin sonuçları elde edilebilir. Ayrıca geliştirilen sistem sadece Twitter üzerinden alınan sosyal medya verilerini kullanarak işlem yapmaktadır. Sisteme başka sosyal medya platformlarından da veri alma yeteneği eklenerek kişisel fikirleri barındıran veri miktarı ve dolayısıyla sistem başarısı artırılabilir.

KAYNAKLAR

[1] Ohmura, A., Kakusho K. ve Okadome T. (2014). "Social Mood Extraction from Twitter Posts with Document Topic Model", International Conference on Information Science and Applications (ICISA), 1-4.

[2] Bollen, J., Mao H. ve Zeng X. (2011). "Twitter Mood Predicts the Stock Market", Journal of Computational Science, 2(1):1-8.

[3] Bouktif S. ve Awad M. A., (2013). "Ant colony based approach to predict stock market movement from mood collected on Twitter", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 837-845.

[4] Qasem M., Thulasiram R. ve Thulasiram P., (2015). "Twitter Sentiment Classification Using Machine Learning Techniques For Stock Markets", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 834-840.

[5] Asur, S. ve Huberman, B. A. (2010). "Predicting The Future With Social Media", International Conference on Web Intelligence and Intelligent Agent Technology, 492–499.

[6] Krushikanth R. A., Merin J., Supreme M., Chan C.-C., Kathy J. L. ve Federico de G., (2013). "Prediction of Movies Box Office Performance Using Social Media", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 1209-1214.

[7] Culotta A. (2010). "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages", Proceedings of the first workshop on social media analytics, 115–122.

[8] Achrekar H., Gandhe A., Lazarus R., Yu S.-H. ve Liu B., (2011). "Predicting Flu Trends Using Twitter Data," IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 702-707.

[9] Soler J. M., Cuartero F. ve Roblizo M., (2012). "Twitter as a Tool for Predicting Election Results", Proceedings of the 2012 IEEE/ACM Inernational Conference on Advances in Social Network Analysis and Monitoring, 1194 – 1200.

[10] Ibrahim M., Abdillah O., Wicaksono A. F. ve Adriani M., (2015). "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results

in a Twitter Nation", IEEE International Conference on Data Mining Workshop (ICDMW), 1348-1353.

[11] Khatua A., Khatua A., Ghosh K. ve Chaki N., (2015). "Can #Twitter_Trends Predict Election Results? Evidence from 2014 Indian General Election", 48th Hawaii International Conference on System Sciences (HICSS), 1676- 1685.

[12] Mahmood T., Iqbal T., Amin F., Lohanna W. ve Mustafa A., (2013). "Mining Twitter Big Data to Predict 2013 Pakistan Election Winner", 16th International Multi Topic Conference (INMIC), 49-54.

[13] Kayahan D., Sergin A. ve Diri B., (2013). "Twitter ile TV Program Reytinglerinin Belirlenmesi", Signal Processing and Communications Applications Conference (SIU), 1-4.

[14] Twitter, Şirket Hakkında, https://about.twitter.com/tr/company/, 01 Şubat 2016.

[15] Twitter API Dokümantasyon, https://dev.twitter.com/rest/public/search, 01 Şubat 2016.

[16] Zemberek-NLP, Doğal Dil İşleme Kütüphanesi, https://github.com/ahmetaa/zemberek-nlp, 01 Şubat 2016.

[17] Weka Veri Madenciliği Yazılımı, http://www.cs.waikato.ac.nz/ml/weka/, 01 Şubat 2016.

[18] Breiman L. ve Cutler A., Random Forest, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, 01 Şubat 2016.

[19] ASP.NET Web API, https://msdn.microsoft.com/library/hh833994.aspx, 01 Şubat 2016.

[20] AngularJS Framework, https://angularjs.org/, 01 Şubat 2016.

[21] CfsSubsetEval Fonksiyonu, http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CfsSubsetEval.html, 01 Şubat 2016.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Cenk AKARSU

Doğum Tarihi ve Yeri : 16.10.1989 İSTANBUL

Yabancı Dili : İngilizce

E-posta : cenk@cenkakarsu.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Bilgisayar Müh.	Yıldız Teknik Üniversitesi	-
Lisans	Bilgisayar Müh.	Trakya Üniversitesi	2012
Lise	Fen Bilimleri	Şişli anadolu Lisesi	2007

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2013 – Devam Ediyor	Garanti Teknoloji	Uzman
2012 - 2013	Etiya Bilgi Teknolojileri	Yazılım Geliştirici
2009 - 2012	Trakya Üniversitesi Bilgi İşlem Daire Başkanlığı	Part Time Yazılım Geliştirici